

AD-A278 776

AL/HR-TP-1994-0003



**ROADMAP: AN AGENDA FOR
JOINT-SERVICE CLASSIFICATION RESEARCH**

**John P. Campbell
Teresa L. Russell
Deirdre J. Knapp**

**Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314**

**DTIC
ELECTE
MAY 02 1994
S G D**

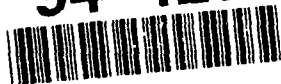
**HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352**

February 1994

Final Technical Paper for Period 1 January 1993 - 30 December 1993

Approved for public release; distribution is unlimited.

94-12979



94 4 28 076

DTIC QUALITY INSPECTED G

**AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS**

**ARMSTRONG
LABORATORY**

NOTICE

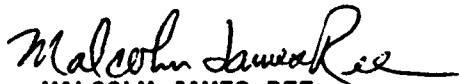
This technical paper is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO) exchange.

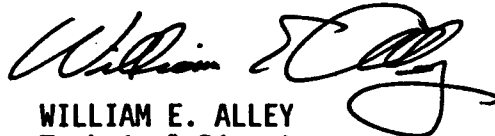
When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.



MALCOLM JAMES REE
Scientific Advisor
Manpower & Personnel Res Div



WILLIAM E. ALLEY
Technical Director
Manpower & Personnel Res Div



WILLARD BEAVERS, Lt Colonel, USAF
Chief, Manpower & Personnel Research Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE February 1994		3. REPORT TYPE AND DATES COVERED Final 1 January 1993 - 30 December 1993	
4. TITLE AND SUBTITLE Roadmap: An Agenda for Joint-Service Classification Research				5. FUNDING NUMBERS C - F33615-91-C-0015 PE - 62205F PR - 7719 TA - 24 WU - 03	
6. AUTHOR(S) John P. Campbell Teresa L. Russell Deirdre J. Knapp					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization (HumRRO) 66 Canal Center Plaza, Suite 400 Alexandria, VA 22314				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Armstrong Laboratory (AFMC) Human Resources Directorate Manpower and Personnel Research Division 7909 Lindbergh Drive Brooks Air Force Base, TX 78235-5352				10. SPONSORING / MONITORING AGENCY REPORT NUMBER AL/HR-TP-1994-0003	
11. SUPPLEMENTARY NOTES Armstrong Laboratory Technical Monitor: Malcolm J. Ree, (210) 536-3942.					
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.				12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) The Joint-Service Classification Research Roadmap is a research agenda designed to enhance the Services' selection and classification research programs. It is composed of numerous research questions that are organized into seven broad activities. Ordered roughly from highest to lowest priority, they are: Building a Joint-Service policy and forecasting data base, capturing criterion policy, modeling classification decisions, developing new job analysis methodologies, investigating fairness issues, conducting criterion measurement research, and conducting predictor-related research. The first two activities, "Building a Joint-Service policy and forecasting data base" and "Capturing criterion policy," will facilitate research planning. "Modeling classification decisions" and "Developing new job analysis methodologies" are activities wherein long-term research is needed. Classification is important because (a) changes in the ASVAB will result in revised composites, (b) recent innovations make classification research timely, and (c) downsizing makes classification more important. Job analysis research is needed to (a) facilitate innovations in predictor and criterion development and (b) facilitate management of selection and classification for future jobs. Fairness is important from a policy perspective. Criterion and predictor-related research are important, but the Services have researched them extensively. Extended research on experimental measures that have yielded promising results is recommended. DTIC QUALITY INSPECTED 3					
14. SUBJECT TERMS Classification Measurement Individual differences Roadmap				15. NUMBER OF PAGES 90	
				16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL		

ROADMAP: AN AGENDA FOR JOINT-SERVICE CLASSIFICATION RESEARCH

PREFACE

Background

The Armstrong Laboratory, the Army Research Institute for the Behavioral and Social Sciences, the Navy Personnel Research and Development Center, and the Center for Naval Analyses are committed to enhancing the overall efficiency of the Services' selection and classification research. This means reducing redundancy of research efforts across Services and improving inter-Service research planning, while ensuring that each Service's priority needs are served. With these goals in mind, the Armstrong Laboratory and the Army Research Institute co-sponsored a project to develop a Joint-Service classification research agenda, or Roadmap. The Roadmap project was performed by the Human Resources Research Organization (HumRRO) for the Armstrong Laboratory, Human Resources Directorate, under Contract No. F33615-91-C-0015, JON 7719 2403. Dr. Malcolm J. Ree of the Armstrong Laboratory was the technical monitor for the contract.

The roadmap project plan had six tasks:

- Task 1. Identify Classification Research Objectives,
- Task 2. Review Classification Tests and Make Recommendations,
- Task 3. Review Job Requirements and Make Recommendations,
- Task 4. Review Criteria and Make Criterion Development
Recommendations,
- Task 5. Review and Recommend Statistical and Validation Methodologies,
- Task 6. Prepare Roadmap.

The first task, Identify Classification Research Objectives, is reported in Russell, Knapp, and Campbell (1992). It involved interviewing selection and classification experts and decision-makers from each Service to determine research objectives. Tasks 2 through 5 were systematic literature reviews of job analytic, predictor development, performance measurement, and methodological research issues. Each review resulted in a report--Russell, Reynolds, and Campbell (1993), Knapp, Russell, and Campbell (1993), Knapp and Campbell (1993), and Campbell (1993)--respectively.

This final report outlines an agenda for future classification research. It attempts to translate information from the literature reviews and information from the Task 1 interviews into a set of critical research needs and describes the kinds of research that have a high potential for meeting those needs. That is, the discussion tries to reflect both the priorities identified in Task 1 and the critical issues in the existing literature that have direct relevance for classification research in the military. The report suggests a possible sequence for the proposed research that takes into account the functional dependencies among a number of the research questions as well as the prerequisite nature of particular kinds of data, to the extent that these things can be known.

The intended audience is composed of the research scientists (primarily psychologists) who are most concerned with planning U.S. military selection and classification research and development. In its current form, the report is not intended for a management audience or an audience that is otherwise unfamiliar with the substantive and methodological issues that are discussed in the previous reports.

It is also the case that the discussion of critical research needs is not in the form of a fully developed Statement of Work or project proposal. A particular need could be addressed in a single project, be made part of a larger project, or spread out over several projects. That is, the separate sections and subsections are not intended to correspond to projects, but to major research issues that should be addressed by whatever project structure is deemed appropriate.

Organization of the Report

The description of future research needs that comprise Roadmap are laid out in the remainder of this report. Section I provides the context for research planning and describes the need to build a Joint-Service policy data base. This data base would contain baseline information such as up-to-date projections of the future size, structure, and missions of the Services. In short, this data base would form a management information system for selection and classification research planners. Section II focuses on job analysis research and explains why job analysis is an important arena for future study. Section III calls for investigations of criterion policy; it shows that the choices that organizations make about criteria (e.g., final school grades, hands-on performance) affect selection and classification systems. Section IV describes criterion measurement research, and Section V centers on predictors and research issues that surround them (e.g., fakability of non-cognitive measures). Section VI suggests research on selection and classification models and the evaluation of specific assignment systems; and Section VII

describes research related to issues of equity/fairness. Throughout Sections I through VII, key research questions are highlighted by italics. Section VIII addresses the sequential properties of the research needs.

The content areas (Sections I through VII) are organized logically, moving from fundamental issues that provide a foundation for a variety of research topics to specific, narrower research content areas. This order does not necessarily reflect importance or priority. There is not a fully determined hierarchy of goals and objectives for classification research, nor is it possible, in this domain of behavioral science research, to specify the information required to achieve a certain goal, or sub-goal completely. Consequently, a research agenda that is fully determined in terms of the specifications for the information that is required and the precise sequence in which it should be obtained is not a realistic aim. However, to the extent that, in the pursuit of specific selection and classification research goals, certain issues must be settled and/or certain kinds of data must be collected before other questions can be addressed, these functional dependencies are identified in the report.

Accession For	
NTIS	CRA&I <input checked="" type="checkbox"/>
DTIC	TAB <input checked="" type="checkbox"/>
Unannounced <input type="checkbox"/>	
Justification _____	
By _____	
Distribution / _____	
Availability Codes	
Dist	Avail and/or Special
A-1	

ROADMAP: AN AGENDA FOR JOINT-SERVICE CLASSIFICATION RESEARCH

EXECUTIVE SUMMARY

The Joint-Service Classification Research Roadmap is a research agenda designed to enhance the efficiency and efficacy of the Services' selection and classification research programs. Another tenet underlying the Roadmap is that the Services must anticipate change and take steps toward meeting future challenges. The uncertainties surrounding today's political and economic environments make innovation and adaptability critical. The Roadmap is intended to facilitate planning and redirection of research efforts.

The Roadmap is composed of a host of research questions that are organized into seven broad activities:

- Building a Joint-Service policy and forecasting data base,
- Developing new job analysis methodologies,
- Capturing criterion policy,
- Conducting criterion measurement research,
- Conducting predictor-related research,
- Modeling classification decisions, and
- Investigating fairness issues.

The following paragraphs provide an overview of each broad activity and discuss the relative priorities of the major research areas.

Building a Joint-Service Policy and Forecasting Data Base

Today's socio-political, technological, and economic environments are in a very dynamic state. The Military Services, like other organizations, must change to meet the demands of these new environments. The parameters that will change during the coming years are the Services' (a) missions, (b) size, (c) structure, (d) technology, and (e) available resources. Building a Joint-Service data base that contains the best possible documentation of forecasts of future missions, resources, and structure is important for preparing the research community to address future needs. This data base and the process involved in creating it would provide important insights for research planning.

Developing New Job Analysis Methodologies

Current job analysis methods are highly appropriate for describing the task content of existing jobs and establishing a link between task performance content and the general cognitive ability determinants of performance on the described tasks. They are less well suited for identifying the knowledge, skill, problem solving strategy, and dispositional determinants of specific performance factors and for anticipating the determinants of performance in future jobs. The gap between what we need to know about jobs and what existing job analysis technologies provide would widen if interpersonal, peer leadership, and communication behaviors become more important. Development of new job analysis methods is needed to provide a more complete specification of performance determinants that will aid in new predictor selection, empirical validation, and synthetic validation, to identify and define expert job performance, and to conduct job analysis in a dynamic environment.

Capturing Criterion Policy

An organization's choice of criteria for personnel research significantly affects research results and, in turn, the design of the selection and classification system. In effect, criterion policy reflects the organization's intended definition for effective performance in that organization, and the types of predictors that are used in selection and classification decision making will depend upon the criteria against which they are judged. Systematic consideration of criterion policy is necessary so that informed decisions can be made about future predictor and criterion development.

Conducting Criterion-Related Research

Over the past decade, the Services have devoted considerable resources to criterion measurement research. The advances in scientific and organizational understanding that have come as a result are many, but much work remains to be done. Outstanding issues are related to (a) determining the impact of using alternative criteria on empirical validation findings, (b) evaluating and improving existing criterion measurement strategies, and (c) developing, as required, innovative measurement strategies to measure nontraditional aspects of performance. Most of these outstanding issues can be at least partially addressed using existing data, thus maximizing the benefit of research dollars already invested by the Services in this arena.

Conducting Predictor-Related Research

The Armed Services Vocational Aptitude Battery (ASVAB) is a highly useful general purpose cognitive predictor. With that in mind, the Services must determine whether it is desirable (i.e., cost effective and consistent with criterion policy) to use other variables such as personality variables in classification equations. If so, two general areas of research questions need attention. First, the Services have made a considerable investment in research on personality, interest, biodata, and psychomotor ability measures. How can the problems associated with their use be overcome? Second, what new predictors will be useful specifically for classification into future military jobs?

Modeling Classification Decisions

The development of the prototypes for the Army's Enlisted Personnel Assignment System (EPAS) and the Air Force's new Processing and Classification of Enlistees (PACE) system, the recent research of Johnson and Zeidner (1990), and the expectation that shrinking resources will make efficient classification even more critical, have heightened interest in further development of classification methodology. There are two principal considerations that in turn lead to a number of specific research questions. (1) Given that there are choices among predictor batteries, performance goals, and criterion measurement methods, how can the maximum potential gain from classification be estimated, and (2) how successfully (i.e., to what degree) can specific operational job assignment procedures capture the potential classification gains?

Investigating Fairness Issues

Fairness issues have important policy implications and are likely to move to the forefront as the demographics of the applicant population change in future decades. Although fairness has been examined and defined repeatedly in the last 30 or 40 years, our understanding of fairness is relatively underdeveloped. The questions that need to be addressed are fundamental ones such as: What is the magnitude of adverse impact or differential prediction for different types of tests? As a whole, how fair are the Services' selection and classification systems?

Classification Research Priorities

Each of the research activities outlined in the Roadmap is important. Even so, priorities do emerge from functional interdependencies among the different areas. The seven activities ordered roughly from highest to lowest priority are:

- Building a Joint-Service policy and forecasting data base,
- Capturing criterion policy,
- Modeling classification decisions,
- Developing new job analysis methodologies,
- Investigating fairness issues,
- Conducting criterion measurement research, and
- Conducting predictor-related research.

The first two activities, "Building a Joint-Service policy and forecasting data base" and "Capturing criterion policy," will facilitate the research planning process. They can be accomplished relatively quickly and should occur soon. "Modeling classification decisions" and "Developing new job analysis methodologies" are activities wherein long-term programmatic research is needed. "Modeling classification decisions" is highly important and time-critical because (a) the Services will need to develop new classification composites soon to accommodate changes in the ASVAB, (b) there have been a number of recent advances in classification research, and (c) downsizing makes effective classification even more important. The importance of job analysis is often understated. Job analysis information might facilitate future innovations in predictor and criterion development, and job analysis methods are needed to help the Services manage selection and classification for future jobs. "Investigating fairness issues" is important from a policy perspective, and several fairness research questions could be addressed immediately. Criterion and predictor-related research are, as always, important, but these areas are ones that the Services have researched extensively. There are research questions in both of these areas that should be addressed soon to extend research on experimental measures that have yielded promising results. Research questions that surround new predictor and criterion measurement will be better informed by accomplishment of some of the other activities.

ROADMAP: AN AGENDA FOR JOINT-SERVICE CLASSIFICATION RESEARCH

Table of Contents

I.	Building a Joint-Service Policy and Forecasting Data Base	1
	The Context of Selection/Classification Research Planning	1
	Military Occupations/Occupational Structures	2
	The Applicant Population	4
	The Mission and Role of the Military Research Community	7
	Development and Codification of Forecasts of the Future	7
	Critical Research Needs	9
II.	Developing New Job Analysis Methodologies	10
	Linking Individual Attributes to Jobs	11
	Cognitive Job Analysis	13
	Future Oriented Job Analyses	14
	Critical Research Questions	16
III.	Capturing Criterion Policy	17
	Critical Research Questions	21
IV.	Conducting Criterion Measurement Research	22
	Impact of Using Alternative Criteria	23
	Utility of Existing Measurement Strategies	24
	Improving Existing Measurement Strategies	25
	Identifying Innovative Measurement Strategies	28
	Critical Research Needs	29
V.	Conducting Predictor-Related Research	31
	Surmounting Obstacles to the Use of High Potential Predictors	34
	Developing New Predictors for Classification	38
	Critical Research Questions	41

VI.	Modeling Classification Decisions	42
	Critical Classification Issues	45
	Evaluation of Specific Assignment Methods	47
	Critical Research Questions	50
VII.	Investigating Fairness Issues	51
	What is Fairness?	51
	Learning More About Adverse Impact and Differential Prediction	53
	Critical Research Questions	56
VIII.	Summary	57
	A Joint-Service Policy and Forecasting Data Base	58
	Criterion Policy	60
	The Classification System Test Bed	60
	Job Analysis Research to Identify Domain Specific Knowledge and Skill Determinants of Job Performance	61
	Job Analysis Methods for Future Jobs	62
	Fairness	63
	Criterion Development	64
	Predictor Development	66
	References	68

List of Figures

Figure 1.	The Joint-Service Classification Research Roadmap	59
-----------	---	----

I. BUILDING A JOINT-SERVICE POLICY AND FORECASTING DATA BASE

Abstract: Today's socio-political, technological, and economic environments are in a very dynamic state. The Military Services, like other organizations, must change to meet the demand of these new environments. The parameters that will change during the coming years are the Services' (a) missions, (b) size, (c) structure, (d) technology, and (e) available resources. Building a Joint-Service data base that contains the best possible documentation of forecasts of future missions, resources, and structure is important for preparing the research community to address future needs. This data base and the process involved in creating it would provide important insights for research planning.

The Context of Selection and Classification Research Planning

Military research has led to significant advances in psychological science for three-quarters of a century. The innovations of World War I and World War II researchers made immeasurable contributions to intelligence testing and selection and classification technology, to name but a few. No other private or public sector organization has matched the accomplishments of the military personnel research community. Today the military laboratories continue to pioneer in many psychological research frontiers (e.g., cognition, learning, abilities measurement, job performance measurement). The Services are unique in their ability to conduct large scale programmatic research that is focused on critical organizational goals.

Continuing the tradition of excellence in military personnel research in the post-Cold War economy will require grappling with a host of new needs. The socio-political-economic environment for military research is dynamic and to some degree unpredictable. Changes are expected in both the supply and demand sides of the classification equation. Changing workforce demographics will substantially affect the

size and nature of the applicant pool. Resource limitations and changing missions will influence the demand side by changing the content and structure of military occupations. All these changes will take place as the research community itself reorganizes and Joint-Service research expands. The dynamics of the military environment are in turn taking place within a very critical time period for the entire national economy. As outlined in the report Workforce 2000 (Johnston & Packer, 1987), the labor force is facing an enormous shortfall in available job knowledge and skill as compared to the demand. Dealing with the shortfall will require a very, very large contribution from applied personnel research. Since the Services have the continued potential for leadership in this domain, the benefits from a massive technology transfer from military research and development to private sector application seem obvious.

Military Occupations/Occupational Structures

The occupational structure of enlisted military jobs has begun to change in response to changing missions and limited resources. For example, in the future, Department of Defense (DoD) involvement in the war on drugs or in the defense of our borders against illegal alien entry may require more small aircraft pilots and small intervention units that operate autonomously. Also, in response to funding limitations, the Services are in the process of redesigning jobs to make the workflow more efficient. It is not possible to know at the moment how far this process will go.

Specialization to generalization. The Services plan to move away from highly specialized jobs as they downsize. Such a move is not without complications. First, there is always resistance to change, and some researchers report that there has been

reluctance in the field to merge jobs. Second, the transition to generalization will require considerable personnel management research effort, such as new job analyses and respecification of job standards. Third, in the past, growth in specialization seems to have been driven by increases in the complexity of technology; and more complex technology means it is harder to design jobs for the generalist. If individuals must be capable of, for example, maintaining several high-technology systems, they will have to be trained on each system. Thus, the military's training investment in each individual is likely to increase with generalization. Also, changes in jobs have implications for future aptitude requirements. As generalization proceeds, enlistees may need to be more versatile and capable of handling diverse tasks.

Team/unit composition. The war on drugs and other mission changes may result in more special operations and low intensity conflicts of short duration. The Services may need the capability to form small, quick-reaction teams of highly specialized personnel for small conflicts around the world. Such emphases imply that the Services may need to expand research on (a) the characteristics individuals need to perform these tasks effectively, and (b) team performance and how individuals contribute to the performance of the team.

Technological advancement. The military is technologically dynamic. For example, advances in shipboard technology have resulted in phasing out the Navy's Boiler Technicians and phasing in a more complex Gas Turbine Technician rating. Such technological changes require retraining personnel on new systems and, as systems are phased out, reallocating individuals across jobs. Thus, it is important to be able to make efficient lateral transfer assignments of personnel.

New systems may also require higher individual technical or cognitive skills, depending upon how "smart" the systems are and how user-friendly maintenance and operations procedures are made. New technology may result in a general shift from concrete observable tasks to cognitively-demanding, non-observable activities (Glaser, Lesgold, & Gott, 1991). In sum, changes in technology may result in more jobs that require advanced technical skills or even jobs that require abilities that are not well measured by current aptitude measures.

The Reserve Components. It is likely that the proportional contribution of the Reserve Components to the total force structure will increase as the proportion for the active duty force decreases. If so, we will need to learn more about the extent to which our knowledge of the active duty force (e.g., skill retention, job performance measurement, selection, and classification) generalizes to the Reserves and what new issues arise in the management of a larger Reserve force.

The Applicant Population

As we approach the year 2000, workforce demographics are changing (Johnston, Faul, Huang, & Packer, 1988). The workforce is aging and will contain proportionally fewer young adults. There will be proportionally more women and minorities, particularly Hispanics. Ree and Earles (1991a) applied demographic trend information to the 1980 Armed Services Vocational Aptitude Battery (ASVAB) norming sample to estimate the effects of demographic change on the Air Force's applicant population. Their findings suggest that the numbers of young people in Armed Services Qualification Test (AFQT) Categories I-IIIa will decrease. There is also concern that substantial

proportions of the workforce will (a) lack the skills and education needed to meet the demands of advanced technological jobs of tomorrow, and (b) lack English language proficiency and/or literacy. So far, reduction in numbers of available youth in the workforce does not appear to have had much impact on the military, probably because reductions in numbers have been matched with reductions in demand, a consequence of the drawdown. However, changes such as the following might be anticipated for training, recruitment, and job assignment functions, as well as for considerations of equity and diversity.

Training. Change in the aptitudes, skill levels, and educational backgrounds of applicants has enormous implications for training. The Services currently recruit and train reasonably high quality enlistees, nearly all high school graduates. Deficiencies in workforce skills could necessitate the development, by DoD, of remedial training programs for basic skills, slower or more detailed technical training, better job aids, or smarter weapons/equipment (that do not require high technical skills to operate or maintain).

Recruitment. Recruiting policies may also need to adapt to changes in the workforce. For example, Navy policy-makers are considering trying to tap the population of AFQT Category I and IIs who are not high school graduates to make better use of the applicant pool. Moreover, compensatory models, that use other variables to compensate for not meeting high school degree requirements, are under development by the Navy. Also, recruiting program planners are considering ways to recruit from different subcultures and deal with language barriers. These efforts can be expected to continue as we move toward a more diverse workforce.

Issues of equity. Issues of equity/fairness in selection and classification testing will probably move toward the forefront as the workforce becomes more diverse. The fairness of levels of representation of minorities in various military occupations will be an issue, especially in times of war or conflict. In short, further research will be needed to find ways to minimize adverse impact and to define and examine fairness within the context of both selection and classification.

Importance of Classification. Operational efficiency and cost-effectiveness, always important for the Services, will be particularly crucial in the post-Cold War economy. Improved classification efficiency is a means by which the aggregate outcome can be increased while holding the selection input constant. Research on classification efficiency, rejuvenated in recent years, has raised several questions that still need answering. For example, are the gains in classification efficiency reported by Johnson and Zeidner (1990, 1991) robust to the application of realistic classification constraints (e.g., aptitude minima) and the Services' current policy of personal choice?

Also, by design or by default, the criteria that an organization chooses to address in its selection and classification system are a reflection of its human resource management goals. However, any set of observed criterion measures is an imperfect reflection of the organization's basic goals (i.e., due to measurement error, deficiency, and contamination) and there are critical issues associated with determining the "fit" between the basic goals and the specific measures used to represent them. The nature of this fit will have a great deal to do with how the operational missions of the research community evolve.

The Mission and Role of the Military Research Community

Finally, in the context of the post-Cold War economy it is likely that the research efforts of the different Services will become more highly integrated. The Services have already taken steps in this direction with the development of Joint-Service Training and Personnel Systems Technology Evaluation and Management (TAPSTEM) committee that monitors selection and classification research. Future moves could accelerate the process.

Development and Codification of Forecasts of the Future

What missions will the future military handle? What will be the future military force structure? Given the dynamic nature of the current environment, the military may undergo vast changes within the next few decades. Yet, predicting the future is a risky business. For organizations, making forecasts about the nature of their future missions and the nature of the constraints that will be imposed on them is even more problematic when their environments are undergoing rapid change. The Services certainly face such a scenario. However, in spite of such complications, and assuming that this kind of exercise has not already been carried out, it would be very useful to develop a management information data base containing information about future missions, future structures, and future constraints to be faced by the Services. Such a data base would serve as a tool for selection and classification investigators who plan research.

While the product--the data base--would be an important resource for decision-makers, the *process* of developing the data base would by itself be informative. Thinking through various scenarios and the impacts on the Services would help planners identify

gaps in planned research, assess priorities, and find novel approaches to solving problems. Because this project would be Joint-Service, the process of developing the data base would also provide a vehicle for clarifying the similarities and differences in the Services' missions and future needs.

The data collection could involve such things as document reviews, interviews with key people, and something akin to "critical incident" workshops in which key participants generate hypothetical critical events that they think have a significant probability of occurrence. The incidents described by the workshop participants should focus on predictions of: (a) specific critical events that could confront one or more of the Services, (b) new constraints that will operate (e.g., budget considerations), and (c) new structures that are deemed necessary (e.g., longer enlistment terms and more generalized technical training).

The outcome of this exercise would be a document, or set of documents, that presents the forecasts in a thorough but user friendly manner for the military personnel research community. If such documentation were current, updatable, and readily available to all researchers, it would go a long way toward maximizing the contributions of the research laboratories to the goals of the DoD and the individual Services and would help prevent research that is obsolete by the time it is finished or research that is difficult to implement or to translate into applied terms. Generating such documentation might be described as organization development or policy research, but it seems to be an important early step in developing and maintaining an effective and achievable selection and classification research agenda during the coming decades.

Critical Research Needs

- Development of policy capturing procedures that are appropriate for eliciting, in the appropriate format, forecasts of future missions, resources, and structures.
- Development of an updatable and user friendly documentation system for the "forecast data base."

II. DEVELOPING NEW JOB ANALYSIS METHODOLOGIES

Abstract: Current job analysis methods are highly appropriate for describing the task content of existing jobs and establishing a link between task performance content and the general cognitive ability determinants of performance on the described tasks. They are less well suited for identifying the knowledge, skill, problem solving strategy, and dispositional determinants of specific performance factors, and for anticipating the determinants of performance in new or future jobs. The gap between what we need to know about jobs and what existing job analysis technologies provide will widen if interpersonal, peer leadership, and communication behaviors become more important. Development of new job analysis methods is needed to provide a more complete specification of performance determinants that will aid in new predictor selection, empirical validation, and synthetic validation, to identify and define expert job performance, and to conduct job analysis in a dynamic environment.

Virtually all personnel research must begin with an appropriate job analysis and the Services have been conducting systematic and thorough job analyses for many years. The Services understand task analysis procedures and their concomitant data analysis techniques very well. There is also a growing literature outside the military that compares the usefulness of alternative job analysis methods for specific purposes (see Cornelius, 1988 and Harvey, 1991). Within the military, both Project A and the Synthetic Validation Project made direct comparisons of alternative methods in terms of their content relevance, reliability, and capability for discriminating among jobs (Campbell & Zook, 1991; Wise, 1991).

Given this rather substantial foundation, what is left to do? Should significant additional resources be directed at job analysis research in the future? The current literature (Knapp, Russell, & Campbell, 1993) says yes. Also, as reflected in the Task 1 results, military researchers expressed interest in novel job analytic approaches, although

job analysis was not considered to be one of the highest priorities (Russell et al., 1992). There are some traditional issues that are still unresolved, as well as new and important issues being raised by the rapidly changing world within which the Services must operate. Within such a context, there seem to be at least three job analytic research needs that warrant attention: (1) developing ways to link individual attributes (or knowledges, skills, and abilities [KSAs]) to jobs efficiently, (2) developing cognitive job analysis methods that focus on the determinants of expert performance, and (3) developing job analysis methods that are responsive to and can anticipate future job change. That is, job analysis methods that will best capture the proposed or forecasted content and structure of future jobs and translate them into forecasts of job requirements.

Linking Individual Attributes to Jobs

What taxonomies of performance determinants (in addition to cognitive ability taxonomies) are useful and valid for linking individual attributes to jobs? While useful technologies exist for describing the content of jobs in terms of either specific tasks or general job behaviors, the development of satisfactory methods for inferring the critical determinants of performance (i.e., individual attributes) in a specific job remains unresolved in the judgment of many military selection and classification experts (Russell et al., 1992). The taxonomies of performance determinants that have received the most attention deal primarily with cognitive and psychomotor abilities, and to a certain extent with personality factors (i.e., the Big Five). As a consequence, knowledgeable judges can, with relatively high agreement, identify the general profile of required abilities for major categories of task content, as reported by Wise, Peterson, Hoffman, Campbell, and

Arabian (1991) in the Phase III report of the Synthetic Validation project. When the objective is to identify the very specific ability determinants for very specific critical tasks, as performed by a very experienced operator, there is little previous research to go on (Ackerman, 1987, 1988). The field has perhaps even less experience with identifying the critical domains of prerequisite experience (e.g., domains of knowledge and skill such as basic mechanics, electronics, and keyboard/computer usage) that determine individual differences in performance in specific jobs.

The research that is needed to address these issues would be fairly long term and programmatic in nature. It took considerable time and empirical research before psychologists could map the general cognitive ability determinants of performance such that expert judgments concerning individual ability attributes could accurately mirror the results of empirical validation. It will also not be a short process for the other domains. However, a focused, programmatic effort to develop taxonomies of critical specific abilities and taxonomies of the critical domains of prerequisite knowledge and skill, such as that already begun in the Learning Abilities Measurement Project (LAMP) should not take quite so long.

There are really two parts to such a program. One concerns the building of valid and useful taxonomies for domains of performance determinants in addition to those that already exist for general cognitive abilities, personality, and interests. The more completely and validly these domains of latent variables can be described and the more knowledgeable subject matter experts can become about them, the more valid will be the linkage of individual attributes to job performance. A major benefit from such a taxonomy would be a much better foundation upon which to base new predictor development and estimates of validity obtained from synthetic validation.

The second part of such a research program would concern the description of job/task content in terms of task components (physical, psychomotor, or cognitive behaviors and processes) that are more basic than surface descriptions of the specific task content, as might be found in the usual narrative job description or task description. "More basic" means that a wide variety of specific tasks could be decomposed into a smaller set of latent processes (e.g., physical, psychomotor, or cognitive behaviors) that underlie individual differences in task performance and which have a more direct link to the attributes. A method of job/task analysis that attempts to do this has been termed "cognitive task analysis," although the same analytic procedures could be used to study the latent processes underlying physical or psychomotor tasks as well as cognitive ones.

Cognitive Job Analysis

Can cognitive job analysis methods be used to efficiently analyze different levels of job performance (e.g., expert, novice) on a large scale? Recent discussions of "cognitive job analyses" (e.g., Glaser et al., 1991) highlight the research findings in expert systems and cognitive psychology that emphasize the differences between the problem solving behavior of experts and the problem solving behavior of novices. In general, experts (i.e., high performers) use strategies that are simpler, more qualitative, and rely less on formal rules than those used by novices. Experts possess a much larger fund of domain specific knowledge and skill. One implication of these findings is that job analytic methods which do not take the desired level of performance into account might miss a good deal of critical information, both in terms of accurately portraying the critical job behaviors that distinguish high from low performance in a specific job and in terms of inferring the

requisite attributes or KSAs. If a goal of classification is to increase the aggregate performance of experienced high level performers then further development of the appropriate job analysis methods for identifying predictors is necessary. Such methods could very well identify KSAs that are different than those which would be identified by conventional procedures. Specific abilities and specific domains of previous experiences could become more important for making differential job assignments.

Current methods of cognitive job analyses used in the development of expert systems are labor intensive and expensive (Glaser et al., 1991). New methods (e.g., variants of the critical incident technique) must be developed that can be used to analyze the fundamental nature of high level performance in a variety of jobs on a larger scale. This research need is shared by the U.S. Department of Labor, which is currently planning for a major revision of the Dictionary of Occupational Titles (Department of Labor, 1992).

Future-Oriented Job Analyses

How can the effects of organizational change on job content and required KSAs be anticipated and described? For the Services, the most critical immediate objective for research and development on job analysis for purposes of serving future classification research is the need to develop methods for analyzing jobs which do not yet exist. Such situations will result from changes in the organization's missions, technology, and structure. How can the effects of such changes on job content and required KSAs be anticipated? To some extent, the relevant issues and some possible procedures are discussed by Harvey (1991). Such methods would be exercises in forecasting using expert

judgment. A number of questions would revolve around the appropriate unit of description, the data collection procedures that would be most useful, and the necessary subject matter expert (SME) qualifications. Methods such as the Nominal Group Technique, the Delphi technique, and content analyses of hypothesized future critical incidents could all prove useful.

It may also be the case that the changes to be experienced by the Services have already been, or are being, experienced by the private sector (e.g., downsizing, a more generalist job structure, increased emphasis on the high-technology work group). In particular, research on job analysis for selection and classification purposes should carefully consider how best to analyze individual performance as it pertains to the role of a work group or team member. What actions constitute that role, what are its necessary KSAs, and what characterizes the high performer versus the low performer?

Given the impending changes in missions, technology, and organizational structure for the Services, the need to anticipate the critical personnel management requirements via future-oriented job analyses seems paramount. This question of how to conduct future oriented job analyses should be addressed immediately. Future criterion and predictor development will depend on using such methods to analyze jobs that may not yet exist.

Critical Research Questions

- What are the specific job analytic methodologies that will best anticipate the performance content and performance determinants for new or future jobs?
- What job analytic procedures would be appropriate for conducting a "cognitive job analysis" of key military entry level, advanced technical, or supervisory/leadership positions?
- Will cognitive job analyses identify critical determinants of high level performance that are not identified by traditional methods?
- For purposes of future predictor research, battery selection, and synthetic estimates of validity, what is the most useful taxonomy of domain specific knowledge and skill, problem solving/trouble shooting strategies, dispositional characteristics (e.g., "personality," "motives," "interests") that can be used to specify a taxonomic model of performance determinants?

III. CAPTURING CRITERION POLICY

Abstract: An organization's choice of criteria for personnel research significantly affects how research results will influence the design of the selection and classification system. In effect, criterion policy reflects the organization's intended definition for effective performance in that organization, and the types of predictors that are used in selection and classification decision making will depend upon the criteria against which they are compared. Systematic consideration of criterion policy is necessary so that informed decisions can be made about future predictor and criterion development.

Policy judgments, whether they are made explicitly or by default, will have considerable influence on the nature of performance criterion development and performance criterion research. For example, the differences represented by the positions of the National Research Council's (NRC) Committee on the Performance of Military Personnel (Wigdor & Green, 1991) and the criterion development work in Project A are considerable. The NRC's position opts for the use of standardized hands-on tests which are based on a randomly selected sample of job tasks. Project A viewed performance as multi-dimensional and not limited to technical task behaviors. The direction taken by criterion development and criterion research will, in turn, have a critical influence on subsequent selection and classification research, whether explicitly recognized or not. For example, will the selection and classification system attempt to identify individuals who would be effective team members and who would demonstrate high levels of peer leadership and support? From the Project A framework, it would. From the NRC point of view, it would not. Some important criterion policy issues facing the Services revolve around how job performance is defined, measured, and used in the selection and classification context.

How should job performance be defined? The job behaviors or actions that are designated as relevant for the organization's goals and the job behaviors or actions that are designated as not relevant for the organization's goals reflect important management decisions. The technical manuals for Army or Marine Corps military occupational specialties, Air Force specialties, or Navy ratings specify the explicit tasks that are goal relevant for each job. However, a job analysis may also identify job behaviors or actions that represent important contributions even though they are not tied to specific task performance. The job analysis might identify relevant behaviors which are not specified in job manuals but which could still be viewed as critically important for effective individual performance, for example, coaching one's peers when they are stuck on a problem, working extra time even when not explicitly asked to do so, or being supportive of other team members. For purposes of selection and classification research, should such actions be counted as critical elements of performance that must be measured? This is a policy decision that should be considered explicitly, or explicitly ignored.

How should job performance be measured? Historically, in the context of the criterion problem, there has been a consistent bias against ratings as a measurement method in favor of "objective" indicators. Certainly there is a case to be made against ratings on the basis of their frequently observed low reliabilities and the high intercorrelations among different dimensions (i.e., halo). However, more recent evaluations have not been so negative (e.g., Campbell, McCloy, Oppler, & Sager, 1993; Nathan & Alexander, 1988). They suggest that ratings are a useful measurement method that deserve additional consideration. As a matter of policy, should the long standing bias against ratings continue, or should the issue be reexamined?

How should performance scores be used in selection and classification research?

Performance is multidimensional and there may be considerable differential prediction (Brogden, 1959) across criterion components within a particular job. That is, scores on the different components of performance within a single job are predicted by different things. This engenders the question of how multiple pieces of information should be combined for decision making purposes, as in making selection decisions and classification decisions. Schmidt and Kaplan (1971) described the issues very clearly. It is a policy decision as to how the multiple pieces of information should be combined. This includes judgments about whether a compensatory or non-compensatory model should be used and what the combinatorial rules (e.g., component weights) should be. In the military these issues come up when deciding what the goals (i.e., predicted scores) for selection versus classification should be. For example, should the goal of classification be to maximize aggregated predicted performance on very broad composites of performance in each job, or should differential job assignments be made on the basis of predicted scores on very specific and very critical performance components that deliberately maximize the distinctions among jobs?

How should important criterion policy decisions like these be approached? One strategy is to ignore them, and for certain purposes this could be a viable strategy. The alternative is to adopt a strategy, or multiple strategies, for considering these policy judgments directly. Some alternatives the Services might consider are the following.

- A Joint-Service working group composed of both senior research and personnel policy experts could be empowered to define and describe the policy issues to be considered. The initial meetings could be conducted using the Nominal Group Technique format (Del Becq, Van de Ven, &

Gustafson, 1975). Consensus on the final set of critical issues could be reached in the usual manner, or by using more formal means of conflict resolution, such as third party consultation. After the issues to be considered have been identified, each Service could develop its position on each issue via some agreed upon strategy of consensus building. For example, within each Service, two rounds of the Delphi technique might be used in which each individual submits a position on each issue and describes the reasons for taking such a position. Consensus building on the policy positions to be advocated could then take place both at the Service and the DoD level.

- An alternative strategy would be to carry out the entire exercise as an organization development effort facilitated by an external party (e.g., using procedures such as described in French, Bell, & Zawacki, 1989; Huse & Cummings, 1985).

An attempt to address these issues should be made as soon as possible, if only to determine that the default position is the most desirable. The judgments that are made will do much to shape the nature of selection and classification research in the coming decades.

Ideally, the DoD research data base would make it possible to evaluate empirically the sensitivity of selection and classification decision making outcomes to alternative criterion measurement strategies. For example, how would using ratings versus standard job sample measures of the same performance factor affect the selection of tests to form a battery and how would it change estimates of gains from classification? Further refinement in the Job Performance Measurement (JPM) project data bank might

make it possible to carry out such sensitivity analyses, at least to a limited degree. That is, to some degree, it is now possible to build a "what if" kind of simulation which allows the decision maker to evaluate the effects of changes in the goals (i.e., the criteria) of selection and classification, changes in the nature of the predictor battery, or changes in the format of the decision making procedure (e.g., one versus two stage) in terms of their predicted effects on selection validity, classification efficiency, total attrition, mean predicted performance, and so forth. In the long run, the more fully the entire system can be modeled, the more useful such sensitivity analyses.

The Services appear to endorse different policies for certain critical aspects of criterion research and criterion measurement (Knapp & Campbell, 1993; Russell, et al., 1992). It would be highly useful to both delineate those policies fully and resolve them as quickly as possible.

Critical Research Questions

- What procedure, or procedures, should be used to facilitate discussions of criterion policy issues?
- What is the nature of the consensus, or lack of consensus, about the goals of criterion measurement in DoD personnel selection and classification research?

IV. CONDUCTING CRITERION MEASUREMENT RESEARCH

Abstract: Over the past decade, the Services have devoted considerable resources to criterion measurement research. The advances in scientific and organizational understanding that have come as a result are many, but much work remains to be done. Outstanding issues are related to (a) determining the impact of using alternative criteria on empirical validation findings, (b) evaluating and improving existing criterion measurement strategies, and (c) developing, as required, innovative measurement strategies to measure nontraditional aspects of performance. Most of these outstanding issues can be at least partially addressed using existing data, thus maximizing the benefit of research dollars already invested by the Services in this arena.

The recently completed Joint-Service Job Performance Measurement (JPM) Project (Harris, 1987) represents the primary foundation for future research efforts in enlisted personnel criterion measurement. JPM criterion measures were developed for 33 military jobs. Hands-on tests were developed for almost all of the jobs; written knowledge tests and rating scales were developed for more than one-half of the jobs. In addition, simulations (e.g., interactive video tests) were developed for a subset of jobs, and archival indices of performance (e.g., training grades) were identified for many of the jobs. Using these instruments, data were collected on over 15,400 enlisted personnel (26,400 if the Army's Project A longitudinal validation sample is included).

Four primary criterion research areas involve:

- (1) Evaluating the impact of alternative criteria on the outcomes of validation studies,
- (2) Evaluating the usefulness of existing criterion measurement strategies,
- (3) Improving the usefulness of existing criterion measurement strategies, and
- (4) Evaluating the need for and identifying innovative criterion measurement strategies.

The ordering of these research areas is a rough reflection of their relative priority. However, prioritization at this level is based more on a rational organization of effort than on the importance of each research area. The first two research areas should be addressed in conjunction with one another. That is, the evaluation of the usefulness of existing measures and the evaluation of the empirical impact of using different criteria for validation research are inextricably tied to each other. Furthermore, before deciding on the resources that should be devoted to improving criterion measurement strategies in the future, it makes sense to evaluate the utility of those measures and the impact of using different criterion measures. If differences across methods are not large, fewer resources should be expended to improve existing methods, and those resources should be directed towards criteria which produce the most meaningful validation results at lowest cost.

Impact of Using Alternative Criteria

Other things being equal, will different criterion measures yield different estimates of validity? Will certain criterion measures yield relatively high estimates of absolute validity and relatively low estimates of differential (i.e., classification) validity or vice versa? To answer these questions, comparisons can be made across different measures of the same latent variable (e.g., job knowledge versus hands-on test measures of technical task performance) or across different latent variables (e.g., technical task performance versus effort and leadership performance). To the fullest extent possible, the effects of reliability differences and method variance should be controlled when such comparisons are made.

These comparisons will help determine the impact of using alternative criteria in selection and classification research. A fuller understanding of this impact is important for at least two reasons. First, future validation studies and related research activities should focus on those latent criterion variables and measures of those variables which hold the most promise for (1) yielding useful validation results and (2) reflecting policy-maker decisions regarding which components of performance are most important for achieving organizational goals. The former characteristic can be evaluated through comparison of empirical results as discussed above, and evaluation of the latter characteristic can be achieved through methods described in Section III. Note that "useful" results provide enough information on which to make decisions regarding operational systems. A second reason for studying the impact of different criterion measures is to permit the choice of how to measure a given latent variable to be determined by empirical findings instead of making the assumption that the most extravagant (i.e., expensive) method is going to yield the most informative results.

Utility of Existing Measurement Strategies

What is the psychometric quality, cost effectiveness, and overall utility of existing measurement strategies taken alone and when used in combination? The reference to existing strategies in this context refers primarily to the measurement techniques used in the Joint-Service JPM project. Much more evaluation of these techniques can be done with the JPM data than we have seen thus far. The assessment of quality should include the criteria discussed in the Task 4 report (i.e., relevance, comprehensiveness, susceptibility to contamination, reliability, discriminability, and practicality) as well as

issues related to measurement bias (see Oppler, Campbell, Pulakos, & Borman, 1992). Evaluation of JPM measures should not be restricted to pre-existing scoring schemes and individual measures. Rather, these activities should include evaluation of alternative scoring strategies, as well as indicators of latent variables that are based on scores from more than one type of measure.

Improving Existing Measurement Strategies

This research area includes two major types of research questions. The first set of questions deals with the sampling of job content for purposes of criterion measurement. The second set of questions pertains to selected criterion measurement methods (i.e., ratings, verbal measures, and administrative indices). Both sets of issues share a common goal of improving existing criterion measurement strategies.

What is the relative validity of various content sampling strategies? For purposes of specifying the content of measures of specific task performance in the JPM project, the Services used a number of different task sampling techniques which varied considerably in complexity. For example, the Marine Corps identified behavioral elements underlying task performance and used this information to avoid selection of tasks which were redundant with regard to these underlying requirements (Felker et al., 1988). In contrast, the Navy experimented with one task selection strategy in which subject matter experts simply recommended a set of tasks for testing (Laabs, Berry, Vineberg, & Zimmerman, 1987). The more complex methods may indeed be more content valid, but is the increase in validity enough to justify the increased time and expense required to use them? The overall results of the validation efforts using the Service's JPM hands-on

work sample tests suggest that the answer to this question may be no. But that conclusion can be tested more systematically, and the real need is to identify the specific procedures which are required to adequately sample job tasks.

To what extent is it better to assess performance comprehensively on a few tasks or to assess performance less thoroughly on a larger proportion of tasks? The answer to this question may depend upon the nature of the job under study, and will probably be at least partially driven by managerial policy. If resources allow, both strategies can be followed. But as criterion measurement resources decrease, a choice between the two strategies is likely to be required for future criterion measure development activities.

What is the best way to sample nontask-based job content? Sampling of nontask-based job content usually comes down to determining the level of specificity at which dimensions of performance will be identified using behaviorally-based job analysis (e.g., critical incident) information. In other words, instead of choosing not to measure some dimensions of performance, researchers generally select a level of dimensional specificity which allows all dimensions to be assessed. The most common application of this strategy is the development of performance rating scales.

The potential for increased use of worker-oriented job analysis strategies, which are by definition not job specific, suggests the need for more attention to behavior (as opposed to task) sampling issues. This type of job analysis information is often used for developing rating instruments, but could also form the foundation for performance and verbal (e.g., job knowledge) tests.

What strategies can be used to maximize the usefulness of ratings data in military research? Recent research findings counter the common tendency to discount the use of ratings as a measurement method. Assuming additional evaluation research supports

this, resources should then be devoted to improving the utility of this measurement method for military criterion assessment purposes. For example, research could be conducted to identify the influences that varying amounts and types of rater training have on the reduction of rater reliance on idiosyncratic theories of performance, reducing halo, increasing rating variability, and so forth. Other research questions concern the amount of rater/ratee familiarity that is required to yield reliable and valid ratings data, and the influence of rater motivation on the quality of data. Simply notifying raters that the data are for-research-only is no guarantee that raters will be motivated to provide accurate data, although it does appear to help.

What types of verbal tests can be used to measure performance factors more comprehensively and accurately? Although these measures do not appear to reflect procedural skill, does their ability to measure declarative knowledge provide a useful contribution to the measure of certain latent criterion variables? Are there ways of developing these tests to maximize this contribution? These research questions can be applied to the various types of verbal tests that were developed in the JPM project (e.g., Situational Judgment Tests, "Walk Through" protocols), as well as to variants of those strategies which have been tried by other organizations (e.g., written essays, various types of situation interview strategies). Given that the primary advantage of this category of tests is efficiency and cost-effectiveness, however, some potential alternatives (e.g., individual interviews) may not be worth extensive consideration.

What can be done to maximize the usefulness of administrative indices of performance? To be useful to researchers, administrative indices of performance must be reliable, valid, comparable across subgroups (e.g., jobs), and permanently recorded in an accessible format (i.e., computerized). Those who devise, administer, use, and/or

maintain operational performance indices also have certain requirements that must be met. Given that there are both operational and research interests in the handling of administrative measures of performance, perhaps this question could best be addressed in a cooperative fashion. For example, researchers could use their expertise in the systematic gathering of information and ideas to get input from each other and from individuals with an operational perspective to help identify administrative performance indices suitable for research and to improve the accuracy and accessibility of archived information. A natural progression from such an effort might be more extensive routine involvement of selection and classification researchers in the development and administration of operational tests (e.g., training tests, specialty knowledge tests).

Identifying Innovative Measurement Strategies

Do we need different content specifications and measurement techniques to assess performance factors anticipated to be important in the future or which are currently important but not usually measured? What will these new measures look like? Latent criterion variables that have not received very much research attention include individual contributions to team performance, good citizenship behaviors, and behaviors and skills associated with successful performance under conditions imposed by operating within foreign cultures. As the Services explore nontraditional types of criterion areas, the need to examine content sampling strategies suitable for use with nontask-based job analyses (described previously) can be expected to increase in importance and complexity. Furthermore, the extent to which existing measurement strategies can adequately measure previously unexplored aspects of performance will need to be examined

systematically. It is reasonable to predict that this examination will point to the need to conduct research on other measurement strategies that might hold out more promise for measuring these latent variables. Innovations might expand upon current measurement strategies which fall within the basic measurement method alternatives (i.e., performance, verbal, ratings, administrative, and direct observation) or they might combine different measurement methods in novel ways. Not only could such innovations allow testing of heretofore unmeasured aspects of performance, but they might also decrease measurement costs and increase the fidelity with which more traditional dimensions of performance are measured. An example of such an innovation might be to obtain more reliable ratings from supervisors by allowing them to observe performance on standardized walk-through exercises. Note that the Air Force used a walk-through performance testing strategy in its JPM effort, but examinees were not evaluated by their supervisors based on their performance on these tests. As another example, an individual's expected responses to various work-related problems might be measured using a combination of supervisor judgments in conjunction with examinee responses to a written or oral situation judgment test covering the same problems.

Critical Research Needs

- Will certain criterion measures yield relatively high estimates of absolute validity and relatively low estimates of differential (i.e., classification) validity and vice versa?

- What is the psychometric quality, cost effectiveness, and overall utility of existing measurement strategies taken alone and when used in combination?
- Can the quality, cost effectiveness, and overall utility of existing measurement strategies be improved by (a) conducting additional research on specific measurement strategies and (b) making job content sampling approaches more efficient and more applicable for nontask-based job content?

V. CONDUCTING PREDICTOR-RELATED RESEARCH

Abstract: The Armed Services Vocational Aptitude Battery (ASVAB) is a highly useful general purpose cognitive predictor. With that in mind, the Services must determine whether it is desirable (i.e., cost effective and consistent with criterion policy) to use other variables such as personality variables in classification equations. If so, two general areas of research questions need attention. First, the Services have made a considerable investment in research on personality, interest, biodata, and psychomotor ability measures. How can the problems associated with their use be overcome? Second and later, what new predictors will be useful specifically for classification into future military jobs?

The Armed Services Vocational Aptitude Battery (ASVAB), the Services' primary selection and classification tool, is a highly useful general purpose cognitive predictor. ASVAB scores (i.e., subtest scores, composites, or the ASVAB general factor scores) are valid predictors of training performance (Earles & Ree, 1992; Ree & Earles, 1991b, 1992; Welsh, Kucinkas, & Curran, 1990) and both first and second-tour job performance (McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Oppler, Peterson, & Russell, in press; Peterson & Rosse, 1992). The ASVAB predicts training success in a host of schools, for a variety of jobs, and in all the Services (Welsh, Kucinkas, & Curran, 1990).

Over the last few years, research has focused on potential modification of the ASVAB. This process began in 1989 when a Technical Advisory Selection Panel (TASP) was established to recommend tests to supplement or enhance the ASVAB. The TASP solicited suggestions for supplemental predictors from the military testing community. After reviewing information about tests, TASP recommended nine tests: three spatial tests, two working memory capacity tests, one figural reasoning test, one perceptual speed test, and two psychomotor tests. These tests formed the Enhanced Computer

Assisted Test (ECAT) battery. Data that have been collected and analyzed on each of the ECAT tests are summarized in Russell et al. (1993).

Currently the ASVAB Review Technical Committee (ART) is assembling information to make decisions about changes in ASVAB content. Changes under consideration involve adding ECAT measures or dropping or modifying some ASVAB subtests. The Navy has overseen the preparation of the ECAT battery and is currently collecting and analyzing data on some of the ECAT measures. All Services are contributing ideas and analyses.

The vehicle for administration of the ASVAB is also under review in a project sponsored by the Defense Manpower Data Center, and known as Concept of Operations (COP). This project generated alternative operational concepts for ASVAB administration. Example concepts involve administering a paper-and-pencil ASVAB but using electronic response pads to collect responses or, alternatively, administering the Computerized Adaptive version of the ASVAB (i.e., CAT-ASVAB). Eight alternative concepts are undergoing cost evaluation.

The ART and COP recommendations for ASVAB content and administration are due in March of 1993. One alternative operational concept will be selected and presented to the Defense Advisory Committee on Military Personnel Tests for review. These decisions will direct predictor research efforts in the near future. Revisions in ASVAB content will necessitate changes in assignment composites. Changes in administration mechanisms will require a good deal of effort and perhaps additional research such as setting up demonstration projects or test sites.

There are two principal areas of predictor research needs: (1) identification/development of predictors that are equitable and (2) identification/development of

predictors that are useful specifically for classification. For the most part, issues of equity are covered in Section VII; this section focuses on the second set of issues.

Given that the ASVAB is a highly useful selection and classification tool, *is it desirable (i.e., cost effective and consistent with criterion policy) to go beyond general cognitive abilities as variables on which to base selection and classification decisions?* The answer must take into account administration costs as well as estimates of the new predictors' incremental validity and classification efficiency. Project A data suggest that the organization's choice of criteria is also key. For example, personality predictors showed significant incremental validity (over and above that of the ASVAB) for predicting performance components such as effort and peer leadership performance and avoiding discipline problems (McHenry et al., 1990). It is not likely that substantial gains in prediction of technical performance criteria will be realized with personality measures. Some but not all of the information needed to answer the cost-effectiveness question is probably available. A concept-of-operations-like project focusing on classification could organize the available information, provide cost estimates, solicit utility judgments, and evaluate alternative scenarios.

Assuming that it would be cost-effective to go beyond general cognitive abilities for classification purposes, future research should focus on: (1) surmounting obstacles to the use of high potential predictors that are in experimental stages, and (2) developing new predictors aimed at classification.

Surmounting Obstacles to the Use of High Potential Predictors

The Services have made a considerable investment in research on personality, interest, biodata, and psychomotor ability tests over the last few decades. Examples of measures that are already supported by a great deal of research are the Assessment of Background and Life Experiences (ABLE), the Vocational Interest Career Examination (VOICE), the Adaptability Screening Profile (ASP), the ECAT target tracking measures, and subtests on the Basic Attributes Test (BAT). In many cases there is strong evidence of the validity (and incremental validity over the ASVAB) of such measures. Even so, there are significant research questions which the Services must deal with before these tests can be implemented. Perhaps the most important of these has to do with personality and other temperament arenas.

Is it possible to develop measures of personality, motives, and values, that are difficult to "fake?" Personality predictors are promising candidates as supplements to the cognitive measures traditionally used by the Services. Recent advances in the area of personality structure have led to new agreement on basic factors around which traits may be organized (Digman, 1990). These factors have helped researchers to be specific about the nature of the criterion relationships that may be expected for personality variables. Meta-analyses have shown personality variables to have consistent useful relationships with a variety of criteria (Tett, Jackson, & Rothstein, 1991). Research indicates that personality measures are valid predictors of certain performance components and contribute unique variance to batteries of cognitive tests, especially for the prediction of "will-do" criteria as well as training attrition (McHenry et al., 1990). Importantly, personality measures appear to show fewer differences between races than do cognitive

measures, and the differences that have been shown tend to favor minority respondents (Reynolds, 1993).

Like personality variables, biodata are effective and valid predictors of a number of important criteria (Reynolds, 1993). Research has indicated that biodata validities can be made generalizable and stable (Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990); thus these measures are worthy of continued consideration as supplements to cognitive predictors of military performance. There is also evidence that biodata may have incremental validity over cognitive measures, especially when predicting non-performance criteria such as attrition (e.g., Trent, in press). Biodata do not yield large differences among the races and evidence of differential validity is slight. Another advantage of biodata is that some measures account for variability in attrition that has traditionally been predicted by educational attainment criteria. Educational credentials have come under fire (cf. Laurence, in press) because they restrict entrance to the military for identifiable groups of individuals (e.g., General Educational Development (GED) recipients). Biodata instruments provide a compensatory measure such that no one particular characteristic would likely exclude an individual. Thus, biodata may face less implementation resistance than other predictors of military adjustment.

The primary issue hindering the actual implementation of personality and biodata measures is the potential for fakability and coachability. It is possible to detect faking with specially designed response scales (e.g., social desirability, unlikely virtues). But in operational settings, suspect responses cannot be simply eliminated from consideration and, to date, no one has identified nor developed a way to adjust or use scores from suspect data. Further research is necessary to determine how to best reduce socially desirable responding. Does one response format work better than another? Does the

vehicle of administration (computer, paper-and-pencil, interview) reduce purposeful faking? Are there other interventions that prevent faking (e.g., giving periodic, tactful feedback on a computer-administered form)? A comprehensive review of the faking and social desirability literature would be a first step in organizing our knowledge in this important area.

Does use of an occupational interest inventory improve classification decisions?

Validation findings indicate that occupational interest measures predict later occupational membership and job satisfaction, however interests do not appear to add much to the prediction of job performance over that accounted for by cognitive and personality predictors. Drawbacks regarding interest measures for selection and classification are: (1) interest measures, like other self-report measures are fakable/coachable, (2) race and gender differences in interests are large, and (3) there is some evidence young adults may lack the maturity and information needed to make stable career choices. In sum, the problems associated with interest measurement are difficult ones. Even so, individuals' job preferences must be taken into account in classification decisions. Currently, the Navy and the Air Force incorporate an occupational preference judgment in their classification algorithms, and recruits from all Services have the option of not enlisting if a preferred job is not available. It would be useful to know whether occupational interest inventories, perhaps empirically keyed, would improve classification over the current methods of obtaining and using occupational preferences.

Are psychomotor abilities useful for classification purposes? Can the influence of practice effects be accounted for? There is reason to expect that psychomotor tests would be useful for classification. However, before implementing psychomotor tests, the Services must deal with the large practice effects associated with them. The effects of

practice on test validity are not well understood. If the test becomes a more reliable measure of the target construct as a result of practice, the increase in reliability should allow for greater validity. Embretson (1987) showed that post-training spatial test scores were more internally consistent and yielded increased predictive validity over pre-training scores. There is, however, some evidence that a test can measure different constructs, depending upon the individual's position on the learning curve (Ackerman, 1987, 1988). Can this aspect of the predictor/criterion latent structure be described? Even large practice effects may have trivial implications for validity, but that is an unanswered research question. A related research question concerns the relationship of psychomotor test validities to the generality/specificity of the performance criterion measure. How specific must the criterion variance be before psychomotor variables maximize their contribution to classification efficiency?

Is there a mechanism for identifying job-relevant physical/psychomotor abilities? Sex differences on psychomotor and physical abilities tests are large (Russell, Tagliareni, & Batley, 1993). As long as these tests are used for classification for combat jobs and combat jobs remain off-limits for women, this is probably a moot point. If, however, combat exclusion policies and laws are modified in the future, a number of issues arise. First, perhaps it will be more important to use psychomotor and physical measures to make classification decisions because a wider range of individuals may be considered for combat jobs. Second, because the sex differences are so large, it will be necessary to show that such tests are based on real job requirements identified through job analyses and shown to be valid predictors of performance. Otherwise, it could be alleged that the Services adopted such tests as a surrogate for combat exclusion policies/laws, since the

use of psychomotor and physical measures could be used to exclude most women from these jobs.

Is administration of physical abilities and psychomotor tests logistically and financially feasible? Another issue is the cost of acquiring special equipment to conduct physical abilities and psychomotor testing. For physical abilities testing, test administrators would also have to be hired and/or trained to validly and reliably measure individuals. In addition, there may well be a space problem to deal with should such testing be implemented at Military Entrance Processing Stations (MEPS). Rooms for testing and space for equipment storage would be needed. Could physical abilities testing occur during or at the end of basic military training instead of at the MEPS?

Developing New Predictors for Classification

Research to identify predictors that demonstrate discriminant validity has not yielded particularly promising results. Usually, differences in validities of ASVAB composites (composed of different subtests) are modest (e.g., Peterson, Gialluca, Borman, Carter, & Rosse, 1990). Johnson and Zeidner (1990, 1991) have reported somewhat larger gains in mean predicted performance using full least squares equations for ASVAB subtests. What types of new predictors might be useful for classification decisions?

Do predictors designed to evaluate existing knowledge and skill in specific domains enhance classification (i.e., discriminant validity, differential validity, mean predicted performance)? Knowledge or achievement measures have been shown to be useful classification tools. For example, Air Force studies in the 1950's showed that

"information" tests were good predictors of pilot and air crew performance and that such tests yielded differential validity (Brogden, 1959) for pilots and navigators (Humphreys, 1986). Recent analyses of the ASVAB have shown that Auto/Shop Information, a knowledge-oriented ASVAB subtest, yields greater discriminant validity than other ASVAB subtests (Oppler, Rosse, Peterson, & Sager, 1993). Information measures tap prior knowledge and experience and the individual's interest in a particular domain, albeit indirectly. Perhaps predictors targeted toward non-technical occupational groups would complement the technical information tests already on the ASVAB (e.g., business information, accounting, computer knowledge).

Use of achievement/information tests has policy implications. Achievement measures are probably more susceptible to opportunity bias than general ability tests. That is, individuals may not have had (or undertaken) the opportunity to acquire specific knowledges. As a result, specific domain tests might yield greater adverse impact against women and minorities than general ability tests. For example, Auto/Shop Information yields the largest sex and race differences compared to other ASVAB subtests (Russell & Tagliareni, 1993). Whether adverse impact should preclude use of the test is a policy issue, not a research issue, but the degree of adverse impact associated with various kinds of tests is a research question. Fairness issues are discussed in greater detail in Section VIII.

Mini-training on critical tasks or in a specific knowledge domain could be used as a part of the testing process to reduce differences attributable to prior experience. For example, the Air Force has developed intelligent tutoring systems for several knowledge domains (e.g., electronics, Pascal programming). Would the post-training test scores be useful for classification purposes?

What variables best predict the team contribution, leadership, communication performance, and citizenship/commitment components of performance that could be more important in the military workforce of the future? The active duty Services of the future will be smaller, with proportionally greater numbers of generalist positions that involve working with technologically sophisticated systems. A greater proportion of military jobs may be in small, quick action, intervention teams that operate with greater independence than present forces. For example, the war on drugs and other mission changes may result in more special operations and low intensity warfare of short duration. The transition to reliance on teams suggests that individual contributions to group performance, peer leadership, citizenship performance, and NCO leadership could be more important performance criteria in the future. What might predict these performance components? Motivation, social intelligence, and values are variables that could be investigated.

Is physical abilities testing needed to classify recruits into physically demanding jobs? The issues that surround physical abilities testing are similar to those of psychomotor testing. Sex differences in physical abilities are enormous, and physical abilities improve with practice. Physical abilities testing requires special equipment and trained scorers. But the Services could administer physical abilities tests during basic military training, after recruits have had the opportunity to improve fitness levels. The key question is whether physical abilities testing is needed to classify recruits into physically demanding jobs.

Critical Research Questions

- Is it desirable (i.e., cost effective and consistent with criterion policy) to go beyond general cognitive abilities as variables on which to base selection and classification decisions?
- Is it possible to develop measures of personality, motives, and values, that are difficult to "fake?"
- Does use of an occupational interest inventory improve classification decisions?
- Are psychomotor abilities useful for classification purposes?
- Are physical abilities measures needed for classification into physically demanding jobs?
- Do predictors designed to evaluate existing knowledge and skill in specific domains enhance classification (i.e., discriminant validity, differential validity, mean predicted performance)?
- What variables best predict the team contribution, leadership, communication performance, and citizenship/commitment components of performance that could be more important in the military workforce of the future?

VI. MODELING CLASSIFICATION DECISIONS

Abstract: The development of the prototypes for the Army's Enlisted Personnel Allocation System (EPAS) and the Air Force's new Processing and Classification of Enlistees (PACE) system, the recent research of Johnson and Zeidner (1990), and the expectation that shrinking resources will make efficient classification even more critical, have all heightened interest in further development of classification methodology. There are two principal considerations that in turn lead to a number of specific research questions. (1) Given that there are choices among predictor batteries, performance goals, and criterion measurement methods, how can the maximum potential gain from classification be estimated, and (2) how successfully (i.e., to what degree) can specific operational job assignment procedures capture the potential classification gains?

Selection and classification experts who were contacted as part of Task 1 of the roadmap project did not ascribe particularly high importance to further research on model development for classification decision making (Russell et al., 1992). However, the current literature has reopened a number of old arguments and has introduced new concerns that highlight several important research issues. Addressing these issues would help create the foundation for a "what if" decision-making test bed that could be used to forecast the effects (on personnel costs, mean predicted performance, etc.) of a wide variety of personnel system changes.

The seminal work of Brogden (1959) formulated the multivariate classification decision as the problem of assigning N people to K jobs, via batch processing, such that every job is filled and some relevant objective function (e.g., aggregate performance) is maximized. In this initial representation of the classification problem, everyone is assigned, there are no quotas or other constraints, the decision is made at the same time

for everybody, and there is one dependent variable to be maximized. It is formally equivalent to the assignment problem in linear programming. Brogden showed that the level of aggregate performance (as well as the mean predicted performance [MPP]) for all people assigned was a function of the absolute level of predictive validity for each job, the level of the intercorrelations among the predicted performance scores for each job, and the number of jobs. Obviously MPP would also be a function of the selection ratio, which would introduce a "not selected" category.

In the classification problem, the predicted scores on each job for each person are taken as a given. As noted by Cronbach and Gleser (1965), computation of the optimal weights for each predictor for each job that maximize the payoff from classification is mathematically intractable. If it were possible to compute them, such weights would be optimal in the sense that they increased absolute validity for each job and decreased the intercorrelations among the predicted scores for each job with the optimal trade off that maximized the gain in predicted performance aggregated over all job assignments.

During the recent past, two concepts have worked to alleviate concern about optimal weights in this context. First, it is widely assumed that if prediction equations were computed so as to maximize predictive validity within each job, then all the differential prediction that it is possible to capture has already been represented. The second concept is the notion that the dominant role of general cognitive ability (g) as a predictor of job performance precludes any significant amount of classification efficiency and makes the utility of classification research approach zero (e.g., Hunter & Schmidt, 1982).

Recently, Johnson and Zeidner (1990) have taken issue with both these notions. Using data from Project A as a representation of a population predictor/criterion

covariance structure for multiple jobs, they have developed a simulation technique that computes the actual gain in mean predicted performance when job assignments are made using (a) various kinds of prediction equations, and/or (b) alternative definitions of job families. Their results suggest that, given current technology, significant gains from classification are certainly possible if predictors can be selected to increase the dimensionality of the predicted score intercorrelation matrix and if job assignments are made to job families that are clustered so as to minimize the intercorrelation of predicted performance scores across clusters and maximize the intercorrelations within clusters. In the Project A data there is indeed a large general factor when the performance component labeled as Core Technical Performance is being predicted, but there is also sufficient specific variance to yield significant gains in MPP from classification. Johnson and Zeidner argue that current assignment systems in the Services do not capture these gains because prediction equations are not developed in the appropriate way and job assignments are made to heterogeneous job families that do not allow the potential benefits from classification to be captured.

Based on the extant literature, including the recent conference on selection and classification issues sponsored by the Army Research Institute in May of 1992, and the nature of the research objectives identified in the Roadmap project (Russell et al., 1992), the research questions of highest priority in this area seem to be as outlined below. They are discussed within two categories. The first deals with basic questions pertaining to general classification issues and the second considers the empirical evaluation of specific differential assignment methods.

Critical Classification Issues

Classification systems can focus on one or more of a variety of goals (Bobko, 1992; Campbell, 1993; Wise, 1992). The design of the system and the way its usefulness is evaluated may be greatly influenced by the goals that are chosen. For example, a system that is solely for the purpose of keeping training seats filled would be designed and evaluated differently than one that attempts to maximize the aggregate level of performance across all jobs at a particular point in time.

A major classification research question concerns how the choice of goal(s), and subsequently the choice of criteria, will influence how predictor information is used and evaluated. That is, *how will goal and criterion choice influence absolute and relative predictor validities and conclusions about gains from classification (e.g., compared to pure selection, current selection procedures, and random assignment)?* For example, consider what happens when either: (a) hands-on performance measures of a representative task samples, (b) hands-on performance measures of only the most critical job tasks, (c) ratings measure of peer leadership and support, or (d) ratings measure of overall performance are used as the function to be maximized. What happens to estimates of absolute validity, discriminate validity, and gains from classification?

Related to the above is the question of whether the existing data banks make it possible to identify goals or performance measures that have the same determinants across jobs (and therefore are best addressed by a selection strategy) and those which do not (and therefore offer potential gains from classification). For example, is it appropriate to select for predicted performance on a composite measure of overall job

performance and then make differential job assignments (i.e., "classify for") on the basis of predicted performance on specific critical tasks within each job?

Another related question concerns how to deal with multiple goals. *Shall the joint effects be evaluated using a compensatory model (e.g., using an additive model with weights provided by subject matter experts), a conjoint model, or some kind of multiple hurdles approach (e.g., applicants can be assigned to a job only if they meet one or more hurdles set for that job).* In sum, to progress further in the development of classification systems it would be very useful to know how inferences about the system are influenced by the goals being addressed.

For predicted task performance, Johnson and Zeidner (1990) have demonstrated that significant classification gains in mean predicted performance are possible, in spite of the large role played by g , if full least square weights are used. Gains are greater to the extent that predictors are sampled from a wider variety of domains and job assignments are made to individual jobs, or to job families that are formed by clustering jobs on the basis of intercorrelations among predicted performance scores. The current operational composites used by the Services do not appear to capture gains from classification. Johnson and Zeidner also challenge the conventional wisdom that classification gains are maximized when predicted validity within each job is maximized. Following Brogden (1959), they make the point that the magnitude of the gain is a joint function of the within-job predictive validity (v) and the intercorrelations among predicted criterion scores (r). Ideally, v is high and r is low. While Johnson and Zeidner acknowledge that the ideal state is never approached they do assert that additional classification gain could

be achieved if test batteries were constructed and prediction equations were generated so as to optimize the tradeoff between increasing v and lowering r .

A critical research question then is, *can methodology be developed to help identify optimal prediction equations in the above sense, that is, prediction equations that lower r while keeping v high?* While a full analytic solution may be mathematically difficult (e.g., see Cronbach and Gleser, 1965) sub-optimal approximations may be possible, or efficient Monte Carlo techniques might be developed, that will yield close approximations to the optimal solution under a range of conditions.

Some of the above issues could be approached by reversing the form of the research question. *That is, instead of asking what gains from classification will result from current predictor and criterion measurement under various conditions, the research objective could be first to specify levels of gain (i.e., from classification over selection) that are important to capture and then determine the nature of the predictor/criterion latent structure that would be required to yield such gains.* For example, what must be the dimensionality of the latent structure, and how reliable must the measures be to achieve an increase in MPP of say .25 standard deviation?

Evaluation of Specific Assignment Methods

At a very applied level, an important set of research questions concerns the evaluation of specific operational (or potentially operational) job assignment methods both in absolute and comparative terms. Operational job assignment systems must deal with a number of constraints. For example, quotas, training seat availability, limited opportunities for batch processing, and cost factors may constrain the optimal solution

such that no operational system can fully capture all the potential classification efficiency available. In absolute terms, classification gains achieved by a specific operational procedure could be compared to the maximum potential gain as estimated by methods such as those used by Johnson and Zeidner (1990, 1991). In relative terms, specific operational assignment procedures could be compared to each other in terms of how well they accomplish a variety of goals such as maximizing aggregate performance, minimizing training costs, and so forth.

Since each of the Services has an operational assignment system and both the Air Force (Pina, Emerson, Leighton, & Cummings, 1988) and the Army (Konieczny, Brown, Hutton, & Stewart, 1990) have experimental systems which are completed but not yet operational, there are a number of specific alternatives that can be compared.

Additional alternatives, in the form of variations of such parameters as (a) the predictors used for selection versus classification, (b) full batch processing for classification versus top down selection (Stage I) plus classification (Stage II), and (c) the introduction of various constraints could also be compared.

Considerations of these kinds of evaluative questions point to the following two major research needs.

First, to evaluate alternative allocation procedures systematically, *can a "test bed" be developed that would simulate the selection and classification procedures of the Army, Air Force, Navy, and Marine Corps, and which would be flexible enough to mimic the structural changes that may occur in the future?* The test bed would incorporate real or hypothesized specifications for both new and existing selection and classification tests, specifications for various alternative criterion measures, and a range of specifications for their latent structures, as well as a range of measurement method characteristics (i.e.,

that could have varying degrees of reliability, validity, and cost). For example, for a particular predictor by criterion latent structure and a particular set of measures of the latent variables, the level of validity, and the maximum potential gains from classification could be computed for some prolonged period of use (i.e., the population value). The question then becomes: For a given set of constraints, what are the effects on costs, the distribution of job assignments, and aggregate criterion scores of using alternative procedures for making selection and job assignment decisions (e.g., compare the new Processing and Classification of Enlistees [PACE] and Enlisted Personnel Allocation System [EPAS])? Also, single stage and two stage models could be compared directly. The current JPM data base should permit construction of a useful test bed that would provide reasonable approximations to operational realities.

Second, if such a test bed were developed, additional questions could be addressed such as *how sensitive are cost factors, the nature of the assignment distributions, and aggregate performance to variations in predictor validities, selection, classification, or performance standards, and assignment priorities (i.e. differential utilities)*? For example, what are the actual effects on overall aggregate performance of raising minimum standards on the predictor(s) versus what are the effects of setting standards on performance and then computing predictor cut scores via regression. "Standards" should be thought of both in terms of critical scores and critical distributions. The sensitivity of system outcomes to a variety of other parameter changes could also be determined. It would give the entire personnel management system of the Services a greatly enhanced "what if" planning and systems design capability.

Critical Research Questions

- How does the choice of the performance construct to be measured (e.g., critical task performance vs. overall performance) and/or the choice of the specific measurement method (e.g., standardized job sample test vs. supervisory ratings) influence estimation of classification gain?
- How can the effects of classification on multiple goals (e.g., increase task performance, increase team contributions, decrease attrition) be evaluated?
- What are the best methods for selecting and weighting predictors so as to optimize classification gains?
- Can the existing data bases and available computer simulation methodologies be used to develop a "test bed" for evaluating the effects of alternative selection and classification procedures?
- Can the test bed be used to evaluate the sensitivity of a particular selection and classification procedure to variations in such parameters as: performance and selection standards, changes in predictor validities, predictor battery dimensionality, selection ratios, etc.?

VII. INVESTIGATING FAIRNESS ISSUES

Abstract: Fairness issues have important policy implications and are likely to move to the forefront as the demographics of the applicant population change in future decades. Although fairness has been examined and defined repeatedly in the last 30 or 40 years, our understanding of fairness is relatively underdeveloped. The questions that need to be addressed are fundamental ones such as: What is the magnitude of adverse impact or differential prediction for different types of tests? As a whole, how fair are the Services' selection and classification systems?

What is Fairness?

Fairness, regardless of context, can be an elusive concept. With regard to employment decision-making, definitions of fairness must take into account societal notions about what is fair, as well as the organization's values and goals. Attempts to define fairness without explicating the value-laden components conflate societal and psychometric goals. Such definitions can appear disingenuous or lack internal consistency. With this in mind, the Society of Industrial and Organizational Psychology (1987) defined fairness as a social rather than a psychometric concept.

Even when designated as a social issue, fairness must be defined operationally to evaluate employee selection procedures. Such attempts have focused on the content, psychometric properties, and use of tests as well as the outcomes of testing. For example, Helms (1992) recently argued that traditional cognitive tests are unfair, based on their content. She asserted that cognitive ability tests measure attributes that are defined by Eurocentric values. Fair tests would measure cognitive attributes defined by other cultures or in the language of other cultures. Her discussion is reminiscent of that which led the attempts to develop culture fair tests in the 1950s, 60s, and 70s. Most of

those studies found that race differences evidenced on so-called culture fair tests were in the same direction as those on traditional measures (see Jensen [1980] for a discussion of culture fair test data). Moreover, definitions of bias in terms of test content have not explained large differences in test score distributions.

Adverse impact occurs when there is "a substantially different rate of selection in hiring, promotion, or other employment decision that works to the disadvantage of members of a race, sex, or ethnic group" (American Institutes for Research, 1992). Adverse impact is not, however, proof of unfairness. Cleary's (1968) psychometric model of fairness is currently accepted by both the Uniform Guidelines (1978) and the Society for Industrial and Organizational Psychology (SIOP, 1987). The Cleary model distinguishes between test bias and test fairness: "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup" (Cleary, 1968, p. 115). In other words, a test is biased when prediction from a common regression equation results in either over- or underprediction of subgroup performance; this is called differential prediction. Overprediction of the performance of a minority group or of women, when a common regression line is used, indicates bias but is generally not considered a fairness problem (SIOP, 1987).

The ASVAB has been evaluated over the years by both military and civilian test experts. According to Eitelberg, Laurence, Waters, and Perelman (1984), the ASVAB predicts success in technical training equitably for males and females and for racial/ethnic groups. The ASVAB typically overpredicts training performance for minority groups (Welsh et al., 1990). In short, the ASVAB meets accepted guidelines for fairness. There is, however, adverse impact in selection and classification (Russell & Tagliareni, 1993),

and although adverse impact does not define fairness, adverse impact certainly raises questions of fairness.

Although fairness has been examined and defined repeatedly in the last 30 or 40 years, our understanding of fairness is relatively underdeveloped. Several basic questions need to be addressed. What is the magnitude of adverse impact in various population subgroups for different types of tests?

Learning More About Adverse Impact and Differential Prediction

What is the magnitude of adverse impact for different variables and for different types of tests? Hopes of finding culture fair cognitive tests, popular in the 1960s and 1970s, were dashed when tests designed to be culturally fair often yielded results favoring whites (Jensen, 1980). Yet, there is meta-analytic evidence that some spatial tests yield smaller sex differences than other tests of the same broad construct (Linn & Petersen, 1985). For example, the ECAT Assembling Objects test has consistently yielded smaller sex differences than other spatial tests (Russell et al., 1993). Additional meta-analyses of sex and race differences in abilities may help shed light on aspects of tests that are related to larger differences.

For all military predictor measures, what is the degree of over- and under-prediction at various score ranges for different performance components and for alternative criterion measures? The body of literature on fairness results is relatively small, probably because fairness analyses require larger samples of minorities than are often available in organizations.

To what extent do over- and under-prediction and adverse impact exist at the latent score level? In a relatively new treatment of fairness, Meredith and Millsap (1992) defined measurement invariance to mean that the same latent variables are measured with the same degree of accuracy in each subpopulation. Gregory (1992) added the principle of structural equity to Meredith and Millsap's definition. According to Gregory, "an unbiased or equitable test is one which, when used to select individuals with the ability to perform a task, yields equal probabilities of selection for all individuals with equal levels of ability relevant to the task, regardless of race, sex, age, etc." (p. 2). In effect, this is the Cleary regression model when the latent variables, rather than the observed predictor and criterion scores, are used on the X and Y axes.

The model uses structural equations modeling to examine the factorial comparability of both the criterion and predictor domains across subgroups. Thus, it can only be tested if multiple predictor and multiple criterion data are available. Ideally, measurement would then proceed using item response theory technology which would in turn require extensive item calibration on large samples and a comparison of item response curves across subgroups. Items would be eliminated if their response curves showed significant differences across subgroups or if differential item functioning (DIF) statistics produced by the Mantel-Haenszel procedure suggest bias (Holland, 1985; Holland & Thayer, 1988). The data demands limit the degree to which the model can be used in research or to guide predictor development. However, it does illustrate a number of fairness issues at their most basic level.

How fair are the Service's selection and classification systems? Fairness is actually broader than just adverse impact or differential prediction on a test. Arvey and Sackett (1993) advocate viewing fairness from a total system perspective. An analysis of this type

is quite appropriate given the complexities of the military selection and classification environment and would probably be more useful to the Services than individual test-based analyses. Unlike the civilian sector, where job applicants typically apply for one specific job, military job applicants are often eligible for more than one job. The military allocates people among jobs. During wartime, charges of unfairness are likely to arise if minorities appear to be disproportionately represented in combat jobs (e.g., Walters, 1991). Similarly, disproportionate numbers of women in administrative and clerical jobs, compared to technical jobs, can appear unfair. Also, some jobs have better advancement opportunities or civilian sector counterparts; underrepresentation of minorities in these jobs is another fairness matter.

For classification systems like those used by the Services, it might be useful to more formally identify and elaborate all the points in the decision system at which disparate treatment could occur (e.g., high school recruitment, applicant screening, classification screening - who qualifies for what jobs or training assignments, retention rates). Does the military inform youth about the kinds of educational experiences that will lead to a preferred job? Does the military seek out well-qualified minority youth? Are selection standards based on job/organizational requirements? What variables enter the classification decision? What are the fairness implications of each step in a Service's classification algorithm? There is a critical need for the development of models which will appropriately represent these issues in context of system fairness. Such model development should benefit from the history of model development for selection fairness. That is, it should make the critical value judgments explicit and should avoid creating representations that are internally inconsistent. If such modeling efforts were successful,

they would enable the Services to pinpoint problem areas and identify ways to promote opportunity while maintaining readiness.

Critical Research Questions

- What is the magnitude of adverse impact for different population subgroups and for different types of tests?
- For all military predictor measures, what is the degree of over- and under-prediction at various score ranges for different performance components and for alternative criterion measures?
- To what extent do over- and under-prediction and adverse impact exist at the latent score level?
- How fair are the Service's selection and classification systems?

VIII. SUMMARY

The classification research roadmap for the Services must deal with at least two major situational forces during the next 10-15 years. First, the Services will undergo considerable change - in mission, structure, size, and technology. Second, the organization and structure of the personnel research function itself will change. Dealing with these uncertainties as best it can must be part of the research community's roadmap. At the same time, everyone must realize that operating in such an environment precludes being able to develop a ten or twenty year plan of research for a specified set of future projects for which all the milestones are already known. Conversely, having to deal with such uncertainties gives specific research and development activities a very high priority.

Given the above context, Roadmap research questions can be placed in a rough order in terms of what should be addressed first, second, third, and so on. There are three basic reasons why an order can be ascribed to a set of research needs. First, a particular research study could be accorded precedence simply because the information is needed to help solve a very important applied problem and the sooner it is obtained the better. Second, addressing a particular research question may be functionally dependent on having answered one or more previous questions. In this case, the questions to be addressed first have a high priority, not because of the intrinsic criticality of their answers, but because they are instrumental for being able to begin work on other issues that have more direct implications for applied problems. Third, a research question may

have little technical value but a very high importance for policy. Such is the case, for example, for fairness issues.

The remainder of this section discusses the ordering of Roadmap research needs. Figure 1 charts this ordering against an illustrative time line.

A Joint-Service Policy and Forecasting Data Base

Effective planning for the future hinges on having available the very best collective judgment about what the future will be like. This makes a Joint-Service Policy and Forecasting Data Base the highest priority. If user friendly documentation of these forecasts is already available then this "need" is moot. If not, then it would have the very highest priority (as shown by Figure 1). Future resources for classification research will be limited and they must be allocated in the best possible way. In large part, this allocation will be made on the basis of forecasts as to what kind of information is needed to maximize the effectiveness of future personnel systems, classification systems in this case.

Capturing and documenting the collective wisdom is not the kind of research that personnel psychologists typically do. However, it is research in terms of the need to develop the methods that capture and synthesize such information in ways that preserve its representativeness and validity.

Assertions that such forecasts, no matter how well done, are always problematic and in some sense always wrong, are well taken. However, this kind of counter argument makes certain implicit assumptions. If this kind of forecast were always wrong, to some

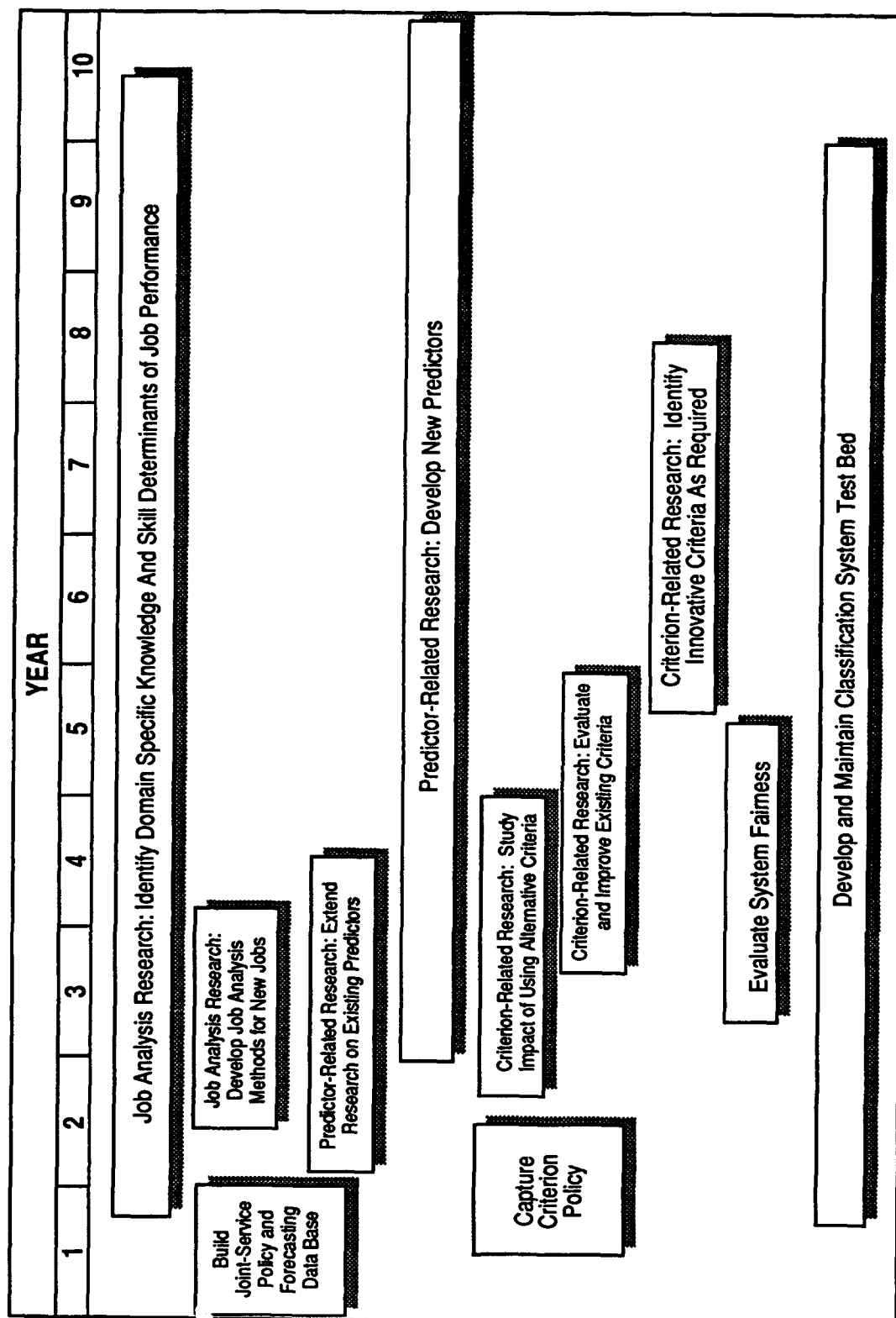


Figure 1. The Joint-Service Classification Research Roadmap

greater or lesser degree, then one implication is that there exists somewhere a method that is error free, or at least significantly less error prone. It would be interesting to know what it is. Second, the counter argument could be based on the assumption that such forecasts are not necessary. After all, to what extent have they been used in the past? However, it seems a truism that anytime research is planned with future application in mind such forecasts are made by somebody either explicitly or by default. In one fashion or another, such forecasts are always used. They should be obtained and documented in the most useful way possible.

Criterion Policy

The second highest priority, which is to some extent contingent on meeting the first priority, is the need to confront the value judgments about criterion measurement that were discussed in Section III. The nature of the research criteria to be used is a direct reflection of the goals of the selection and classification system. A criterion "theory" cannot be an afterthought. In a very real sense, it drives everything else. Using "available" measures simply substitutes a default model.

The Classification System Test Bed

As it was briefly described in Section VII, a functional test bed that simulates the major features of the military selection and classification system would be a valuable means by which the effects of alternative predictor batteries, alternative decision models, and variations in constraints could be evaluated. For example, it would allow researchers to estimate the sensitivity of mean predicted performance (MPP) to changes in

constraints and to changes in the latent structure of the predictor-criterion space. Also, this is one major way in which the implications of small or large departures from a unidimensional view (i.e., g in the ability domain) could be estimated. Other questions, such as how drastically certain constraints can limit the degree to which potential gains in MPP can be realized, or how changes in the decision making system would affect costs could also be investigated. Such a test bed would help do for classification evaluation what the developments in modeling operational utility have done for the evaluation of single stage selection procedures (e.g., see Boudreau, 1991).

As shown in Figure 1, development of a classification system test bed would be a long-term research effort that would begin soon.

Job Analysis Research to Identify Domain Specific Knowledge and Skill Determinants of Job Performance

While useful technologies exist for describing the content of jobs in terms of either specific tasks or general job behaviors, the development of satisfactory methods for inferring the critical determinants of performance (i.e., individual attributes) in a specific job remains unresolved in the judgment of many military selection and classification experts (Russell et al., 1992). The taxonomies of performance determinants that have received the most attention deal primarily with cognitive and psychomotor abilities, and to a certain extent with personality factors (i.e., the Big Five). As a consequence, knowledgeable judges can, with relatively high agreement, identify the general profile of required abilities for major categories of task content, as reported by Wise, Peterson, Hoffman, Campbell, and Arabian (1991) in the Phase III report of the Synthetic

Validation project. When the objective is to identify the very specific ability determinants for very specific critical tasks, as performed by a very experienced operator, there is little previous research as a guide (Ackerman, 1987, 1988). The field has perhaps even less experience with identifying the critical domains of prerequisite experience (e.g., domains of knowledge and skill such as basic mechanics, electronics, and keyboard/computer usage) that determine individual differences in performance in specific jobs.

The research that is needed to address these issues would be fairly long term and programmatic in nature as shown in Figure 1. It took a considerable length of time and a lot of empirical research before psychologists could map the general cognitive ability determinants of performance such that expert judgments concerning individual ability attributes could accurately mirror the results of empirical validation. It will also not be a short process for the other domains. However, a programmatic and focused effort to develop taxonomies of critical specific abilities and taxonomies of the critical domains of prerequisite knowledge and skill, such as that already begun in the Learning Abilities Measurement Project (LAMP) should not take quite so long.

Job Analysis Methods for Future Jobs

Current methods of job analysis emphasize almost exclusively the procedures to be used for studying already existing jobs. In that sense they are very static in nature. However, a potentially critical problem is how to design a selection system that can be used to select people for new jobs--jobs that do not yet have any job incumbents. Designing training programs that will be ready when new jobs come on line is part of the same issue. Unless the results of addressing priority number one suggest otherwise, a

major problem for military selection and classification systems in the future will be to anticipate how to make job assignments to positions which do not yet exist, because of approaching technological changes, changes in the missions to be accomplished, or changes in the structure of the way jobs are organized. Since the design of selection and classification and training systems depend on job analysis information, it would be very useful, if not absolutely necessary, to develop job analysis methods which can make the best possible use of available information to describe what the future positions will be like.

The research needed to begin development of these kinds of job analytic methods can begin very soon. Its later stages could be informed by the information obtained as part of priority one and priority two, but the planning for the research, developing statements of work, and initial model development could begin right away.

Fairness

Fairness is a very important policy issue for the Services. Unlike the civilian sector, where job applicants typically apply for one specific job, military job applicants are often eligible for more than one job. The military allocates people among jobs to a degree. During wartime, charges of unfairness are likely to arise if minorities appear to be disproportionately represented in combat jobs (e.g., Walters, 1991). Similarly, disproportionate numbers of women in administrative and clerical jobs, compared to technical jobs, can appear unfair. Also, some jobs have better advancement opportunities or civilian sector counterparts; underrepresentation of minorities in these jobs is another fairness matter.

Although fairness has been examined and defined repeatedly in the last 30 or 40 years, our understanding of fairness is relatively underdeveloped. Several basic questions need to be addressed. What is the magnitude of adverse impact for different types of tests? How fair are the Services' selection systems?

Criterion Development

Future criterion development efforts should be guided by criterion policy and by a more complete understanding of the impact of using different types of criteria on empirical validation findings. This direction can be provided by the "Capture criterion policy" activities previously described and by analyzing existing Joint-Service JPM project data to examine more thoroughly the empirical impact of alternative criterion measures. For example, do the Services, by design, wish to make "critical job tasks" and job sample tests the construct and measurement method of choice? If they do, how will this effect estimates of selection and classification validity compared to such estimates obtained from training criteria? Such a research direction is the first of three research areas directly related to criterion-related research which are identified in Figure 1.

In terms of the resources that were devoted to it, the JPM Project may never be repeated. Full blown job sample measurement is labor intensive, time consuming, and expensive. Based on available information, Section IV identified three measurement methods that might prove to be less expensive and easier to use and that should perhaps receive the bulk of the research attention. These were: (1) variations of the walk through procedure that do not require expensive and time consuming set ups; (2) rating methods that are based on appropriate job analyses, identification of the most

knowledgeable raters, and utilization of effective rater training techniques; and (3) development of an administrative record data collection system that could be made operational, would be useful for ongoing performance appraisal, and would also be useful as a performance criterion for research purposes. Further development of such measures is necessary if the classification validity/efficiency of new selection and classification tests is to be evaluated on a comprehensive scale. To a considerable degree, the nature of this criterion development work will be driven by policy decisions regarding the criterion models of choice, and thus is functionally dependent on them. It is also true that such a measurement technology should be developed, evaluated, and decided upon before the results of new predictor development are to be evaluated. Evaluation of new predictor innovations will be dependent on having the appropriate criterion model and measurement technology in place. Using inappropriate criterion measures to evaluate new predictors could produce very misleading results. Thus, the second criterion-related research area identified in Figure 1 is "evaluate and improve existing criteria."

The last criterion-related research area shown in Figure 1 is "Identify innovative criteria as required." This research area follows the onset of all of the others identified in Figure 1. Its positioning reflects the fact that the need for criterion which measure nontraditional aspects of job performance, and the nature of those new criteria, will unfold as the other research activities (e.g., new job analysis methods) are conducted and the military work environment continues to evolve.

Predictor Development

The effective planning of predictor research is dependent to some degree on the choices that are made regarding criterion models. If classification research focuses only on overall measures of global job performance, measures of *g* will continue to dominate the selection and classification test battery. If classification research focuses also on the specific components of performance that maximize the differences across jobs development of additional predictor variables would be potentially more useful. To the extent that a multi-dimensional perspective of performance is preferred, two areas of research would be very important for the Services.

First, several experimental predictors have yielded promising results, but additional research is needed before they can be implemented. Extending research on these promising predictors is important (see Figure 1). For example, there is strong evidence that non-cognitive (e.g., personality) measures provide incremental validity (over that afforded by the ASVAB) for the prediction of certain kinds of criteria (e.g., attrition, peer leadership, effort). It is well-known, though, that fakability/coachability is a significant barrier to the use of personality measures and practice effects inhibit use of psychomotor and other tests. Finding ways to overcome these problems is important.

Second, once job analysis and criterion definition research projects make headway, the focus of S&C research may shift to predictor development, particularly toward the development of predictors likely to be useful for classification purposes. As indicated by the past and current research, the types of predictors that seem to hold the most promise are certain specialized abilities and domain specific areas of knowledge and skill. The implications of the model of performance determinants developed in the LAMP Project,

the validities of the Auto-Shop and Electronics subtests of the ASVAB, and the research of Ackerman (1987, 1988) and others all point in this direction. However, perhaps more than any other area of research discussed in this report, research on new selection and classification measures is dependent upon forecasts of future organizational missions, on job analyses of the new jobs that will result from restructuring or from new technology, and on the choices that are made regarding criteria and selection and classification goals.

REFERENCES

- Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. Psychological Bulletin, 102, 3-27.
- Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. Journal of Experimental Psychology: General, 117, 288-318.
- American Institutes for Research (1992). A guide to test validation. Washington, DC: The ERIC Clearinghouse on Tests, Measurement, and Evaluation.
- Adler, T. (1992). Labor Department hires APA to help revise job title book. APA Monitor, 23 (11), 16.
- Arvey, R. D., & Sackett, P. R. (1993). Fairness in selection: Current developments and perspectives. In N. Schmitt, W. C. Borman and Assoc. (Eds.) Personnel selection in organizations. San Francisco: Jossey Bass.
- Bobko, P. (1992). Issues in operational selection and classification systems: Comments and commonalities. Proceedings of the Army Research Institute Selection and Classification Conference. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Boudreau, J. W. (1991). Utility analysis for decisions in human resource management. In M. Dunnette & L. Hough (Eds.) Handbook of I/O Psychology (Rev. Ed. - Vol. II). Palo Alto: Consulting Psychologists Press, 621-745.
- Brogden, H. E. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. Educational and Psychological Measurement, 19, 181-190.
- Campbell, J. P. (Ed.) (1993). Building a joint-service classification research roadmap: Methodological issues in selection and classification. Alexandria, VA: Human Resources Research Organization.
- Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt, W. C. Borman and Assoc. (Eds.) Personnel selection in organizations. San Francisco: Jossey Bass.
- Campbell, J. P., & Zook, L. M. (Eds.) (1991). Improving the selection, classification, and utilization of Army enlisted personnel: Final report on Project A (ARI RR 1597). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

- Cleary, T.A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.
- Cornelius, E. T. III (1988). Practical findings in job analysis research. In S. Gael (Ed.) The job analysis handbook for business, industry, and government. New York: Wiley, Vol I, 48-68.
- Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions (2nd ed.). Urbana, IL: U. of Illinois Press.
- Del Becq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). Group techniques for program planning: A guide to nominal group and delphi processes. Glenview, IL: Scott Foresman.
- Department of Labor, U.S. Employment Service (1992). Interim Report from the Advisory Panel for the Dictionary of Occupational Titles. Washington, D.C.: Author.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. Annual Review of Psychology, 41, 417-440.
- Earles, J. A., & Ree, M. J. (1992). The predictive validity of the ASVAB for training grades. Educational and Psychological Measurement, 52, 721-725.
- Eitelberg, M. J., Laurence, J. H., Waters, B. K., & Perelman, L. S. (1984). Screening for service: Aptitude and education criteria for military entry. Washington, DC: Office of the Assistant Secretary of Defense.
- Embretson, S. E. (1987). Improving the measurement of spatial aptitude by dynamic testing. Intelligence, 11, 333-358.
- Felker, D. B., Crafts, J. L., Rose, A. M., Harnest, C. W., Edwards, D. S., Bowler, E. C., Rivkin, D. W., & McHenry, J. J. (1988). Developing job performance tests for the United States Marine Corps infantry occupational field (AIR-47500-FR 9/88). Washington, D.C.: American Institutes for Research.
- Fleishman, E. A., & Hempel, W. E. (1954). Changes in factor structure of a complex psychomotor test as a function of practice. Psychometrika, 19, 239-252.
- French, W. L., Bell, C. H., & Zawacki, R. A. (1989). Organizational development: Theory, practice, & research (3rd ed.) Homewood, IL: BPI/Irwin.
- Glaser, R., Lesgold, A., & Gott, S. (1991). Implications of cognitive psychology for measuring job performance. In Performance assessment for the workplace: Volume II technical issues. Eds. Wigdor, A.K., & Green, B.F. Jr. Washington DC, National Academy Press.

- Gregory, K.L. (1992). A reconsideration of bias in employment testing from the perspective of factorial invariance. Doctoral dissertation, University of California at Berkeley.
- Harris, D. A. (1987, March). Job performance measurement and the Joint-Service project: An overview. In Proceedings of the Department of Defense/Educational Testing Service Conference on Job Performance Measurement Technologies. San Diego, CA.
- Harris, D. A., McCloy, R. A., Dempsey, J. R., Roth, C. Sackett, P. R., Hedges, L. V., Smith, D. A., & Hogan, P. F. (1991). Determining the relationship between recruit characteristics and job performance: A methodology and a model. (FR-PRD-90-17). Alexandria, VA: Human Resources Research Organization.
- Harvey, R. J. (1991). Job analysis. In M. Dunnette & L. Hough (Eds.) Handbook of I/O Psychology (Rev. ed. - Vol. II). Palo Alto: Consulting Psychologists Press 71-163.
- Helms, J.E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? American Psychologist, 47, 1083-1101.
- Holland, P. W. (1985). On the study of Differential Item Performance without IRT. Proceedings of the Military Testing Association, October.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Brown (Eds.). Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Horn, J. L. (1989). Cognitive diversity: A framework of learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), Learning and individual differences. New York: Freeman, 61-116.
- Humphreys, L. G. (1986). Commentary. Journal of Vocational Behavior, 29, 421-437.
- Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M. D. Dunnette & E. A. Fleishman (Eds.) Human performance and productivity: Human capability assessment (Vol. 3). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Huse, E. F., & Cummings T. G. (1985). Organizational development and change (3rd Ed.). St. Paul, MN.: West.
- Jensen, A. R. (1980). Bias in mental testing. New York: Free Press.
- Johnson, C. D., & Zeidner, J. (1990). Classification utility: Measuring and improving benefits in matching personnel to jobs (IDA Paper P-2240). Alexandria, VA: Institute for Defense Analysis.

- Johnson, C. D., & Zeidner, J. (1991). The economic benefits of predicting job performance: Vol. II classification efficiency. New York: Praeger.
- Johnston, W. B., & Packer, A. E. (1987). Workforce 2000. Indianapolis, IN: Hudson Institute.
- Johnston, W. B., Faul, S., Huang, B., & Packer, A. H. (1988). Civil Service 2000. Washington, DC: US Government Printing Office.
- Knapp, D. K., & Campbell, J. P. (1993). Building a Joint-Service classification research roadmap: Criterion-related issues. Alexandria, VA: Human Resources Research Organization.
- Knapp, D. K., Russell, T. R., & Campbell, J. P. (1993). Building a Joint-Service classification research roadmap: Job analysis methodologies. Alexandria, VA: Human Resources Research Organization.
- Konieczny, F. B., Brown, G. N., Hutton, J., & Stewart, J. E. (1990). Enlisted personnel allocation system: Final report. (ARI Technical Report 902). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Laabs, G. J., Berry, V. M., Vineberg, R., & Zimmerman, R. (1987, March). Comparing different procedures of task selection. In Proceedings of the Department of Defense/Educational Testing Conference on Job Performance Measurement Technologies. San Diego, CA.
- Laurence, J. H. (1993). Education standards and military selection: From the beginning. In J. H. Laurence & T. Trent (Eds.), Adaptability screening for the military. Washington, DC: Department of Defense.
- Linn, M. C., & Peterson, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. Child Development, 56, 1479-1498.
- McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. Personnel Psychology, 43, 335-354.
- Meredith, W. & Millsap, R.E. (1992). On the misuse of manifest variables in the detection of measurement bias. Psychometrika, 57, 289-311.
- Nathan, B. R., & Alexander, R. A. (1988). A comparison of criteria for test validation: A meta analytic investigation. Personnel Psychology, 41, 517-536.
- Oppler, S. H., Campbell, J. P., Pulakos, E. D., & Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results, and conclusions. Journal of Applied Psychology, 77, 201-217.

- Oppler, S. H., Peterson, N. G., & Russell, T. L. (in press). Basic LVI validation results. In J. P. Campbell and L. Zook (Eds.), Building and retaining the career forces: FY 1991 annual report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Oppler, S. H., Rosse, R. L., Peterson, N. G., & Sager, C. (1993, February). A presentation to the ASVAB Technical Review Committee. San Diego, CA.
- Peterson, N. G., Gialluca, K. A., Borman, W. C., Carter, G. W., & Rosse, R. L. (1990). An investigation of methods for simplifying Navy classification (Institute Report 189). Minneapolis, MN: Personnel Decisions Research Institute.
- Peterson, N. G., & Rosse, R. L. (1992, August). Prediction of later career performance with entry-level performance. Paper presented at the Annual Convention of the American Psychological Association, Washington, D.C.
- Pina, M., Jr., Emerson, M. S., Leighton, D. L., & Cummings, W. (1988). Processing and Classification of Enlistees (PACE) system payoff algorithm development. (AFHRL-TP-87-41). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Ree, M. J., & Earles, J. A. (1991a). Estimates of available aptitude as a consequence of demographic change (AL-TP-1991-0019). Brooks, AFB, TX: Armstrong Laboratory.
- Ree, M. J., & Earles, J. A. (1991b). Predicting training success: Not much more than g. Personnel Psychology, 44, 321-332.
- Ree, M. J., & Earles, J. A. (1992). Subtest and composite validity of ASVAB forms 11, 12, and 13 for technical training courses (AL-TR-1991-0107). Brooks AFB, TX: Armstrong Laboratory.
- Reynolds, D. H. (1993). Personality, interest, and biographical attribute measures. In T. L. Russell, D. H. Reynolds, & J. P. Campbell (Eds.). Building a Joint-Service classification research roadmap: Individual differences measurement. Alexandria, VA: Human Resources Research Organization.
- Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? Journal of Applied Psychology, 75, 175-184.
- Russell, T. L., Knapp, D. J., & Campbell, J. P. (1992). Building a Joint-Service classification research roadmap: Defining research objectives (HumRRO IR-PRD-92-10). Alexandria, VA: Human Resources Research Organization.

- Russell, T. L., Reynolds, D. H., & Campbell, J. P. (Eds.) (1993). Building a Joint-Service classification research roadmap: Individual differences measurement. Alexandria, VA: Human Resources Research Organization.
- Russell, T. L., & Tagliareni, F. A. (1993). Operational predictors. In T. L. Russell, D. H. Reynolds, & J. P. Campbell (Eds.). Building a Joint-Service classification research roadmap: Individual differences measurement. Alexandria, VA: Human Resources Research Organization.
- Russell, T. L., Tagliareni, F. A., & Batley, L. (1993). Cognitive, psychomotor, and physical attribute measures. In T. L. Russell, D. H. Reynolds, & J. P. Campbell (Eds.). Building a Joint-Service classification research roadmap: Individual differences measurement. Alexandria, VA: Human Resources Research Organization.
- Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. Personnel Psychology, 24, 419-434.
- Society for Industrial and Organizational Psychology, Inc. (1987). Principles for the Validation and Use of Personnel Selection Procedures. (Third Edition) College Park, MD: Author.
- Tett, B. P., Jackson, D. N., & Rothstein, M. R. (1991). Personality measures as predictors of job performance: A meta-analytic review. Personnel Psychology, 44, 703-742.
- Trent, T. (1993). The Armed Services Applicant Profile (ASAP). In J. H. Laurence & T. Trent (Eds.), Adaptability screening for the military. Washington, DC: Department of Defense.
- Uniform Guidelines in Employee Selection Procedures (1978). Federal Register, 43, 38290-38315.
- Waters, B. (1992). President's message. The military psychologist: The official newsletter of Division 19 of the APA, 9, 1.
- Walters, R. (1991, March). African-American participation in the All Volunteer Force: Lessons from the Persian Gulf Crisis. Testimony before the U.S. House of Representatives Committee on Armed Services.
- Welsh, J. R., Jr., Kucinkas, S. K., & Curran, L. T. (1990). Armed Services Vocational Aptitude Battery (ASVAB): Integrative review of validity studies (AFHRL-TR-90-22). Brooks AFB, TX: U.S. Air Force Human Resources Laboratory.
- Wigdor, A. K., & Green, B. F. (Eds.) (1991). Performance assessment for the Workplace: Report of the Committee on the Performance of Military Personnel. National Research Council. Washington, DC: National Academy Press.

- Wise, L. L. (1992). Goals of the selection and classification decision. Proceedings of the Army Research Institute's Selection and Classification Conference. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Wise, L. L., Peterson, N. G., Hoffman, R. G., Campbell, J. P., & Arabian, J. M. (1991). Army synthetic validity project report of phase III results, volume I (Report 922). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.